1-1-2008

A Bayesian testlet response model with covariates : a simulation study and two applications.

Su G. Baldwin
University of Massachusetts Amherst

A BAYESIAN TESTLET RESPONSE MODEL WITH COVARIATES: A SIMULATION STUDY AND TWO APPLICATIONS

A Dissertation Presented

by

SU G. BALDWIN

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF EDUCATION

February 2008

School of Education
Educational Policy Research and Administration
Research and Evaluation Methods Program

A BAYESIAN TESTLET RESPONSE MODEL WITH COVARIATES: A SIMULATION STUDY AND TWO APPLICATIONS

A Dissertation Presented

by

SU G. BALDWIN

Approved as to style and content by:

Lisa A. Keller, Chair

Ronald K. Hambleton, Member

Erin M. Conlon, Member

Christine B. McCormick, Dean
School of Education

A BAYESIAN TESTLET RESPONSE MODEL WITH COVARIATES: A SIMULATION STUDY AND TWO APPLICATIONS

A Dissertation Presented

by

SU G. BALDWIN

Approved as to style and content by:

Lisa A. Keller, Chair

Ronald K. Hambleton, Member

Erin M. Conlon, Member

Christine B. McCormick, Dean
School of Education

DEDICATION

To my husband Peter.

ACKNOWLEDGMENTS

Foremost, I would like to thank my committee chair and friend Lisa A. Keller for all her support, guidance, and encouragement. I am very lucky to have known her both professionally and personally. I would also like to express my deepest gratitude to my committee member Ronald K. Hambleton. His contribution to my education has been invaluable in every possible way. As a member of my committee, I am indebted to Erin M. Conlon for her kindness and contribution toward completion of this dissertation. In addition, I would like to thank Steve G. Sireci and Craig S. Wells for their unconditional support, enthusiasm, and guidance throughout my education.

I would also like to thank my friends and colleagues at Hills South. Special mention goes to Rob Keller, Ana Karantonis, and Stephen Jirka who made psychometrics "fun" for me.

I would be remiss not to acknowledge the excellent mentorship Howard Wainer and Eric T. Bradlow have provided me throughout the process of writing this dissertation. Without their guidance, I would be long lost in the jungle of parameter space. I would also like to extend my greatest appreciation to the National Board of Medical Examiners, particularly Brian E. Clauser, for the generous financial support given to this project.

Lastly, and most importantly, I would like to thank my husband Peter. Without him, none of this would be possible.

ABSTRACT

A BAYESIAN TESTLET RESPONSE MODEL WITH COVARIATES: A SIMULATION STUDY AND TWO APPLICATIONS

FEBRUARY 2008

SU G. BALDWIN, B.S., MIDDLE EAST TECHNICAL UNIVERSITY, ANKARA, TURKEY

M.A., BOGAZICI UNIVERSITY, ISTANBUL, TURKEY

Ed.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Lisa A. Keller

Understanding the relationship between person, item, and testlet covariates and person, item, and testlet parameters may offer considerable benefits to both test development and test validation efforts. The Bayesian TRT models proposed by Wainer, Bradlow, and Wang (2007) offer a unified structure within which model parameters may be estimated simultaneously with model parameter covariates. This unified approach represents an important advantage of these models: theoretically correct modeling of the relationship between covariates and their respective model parameters. Analogous analyses can be performed via conventional post-hoc regression methods, however, the fully Bayesian framework offers an important advantage over the conventional post-hoc methods by reflecting the uncertainty of the model parameters when estimating their relationship to covariates.

The purpose of this study was twofold. First was to conduct a basic simulation study to investigate the accuracy and effectiveness of the Bayesian TRT approach in

estimating the relationship of covariates to their respective model parameters. Additionally, the Bayesian TRT results were compared to post-hoc regression results, where the dependent variable was the point estimate of the model parameter of interest. Second, an empirical study applied the Bayesian TRT model to two real data sets: the Step 3 component of the United States Medical Licensing Examination (USMLE™), and the Posttraumatic Growth Inventory (PTGI) by Tedeschi and Calhoun (1996).

The findings of both simulation and empirical studies suggest that the Bayesian TRT performs very similarly to the post-hoc approach. Detailed discussion is provided and potential future studies are suggested in chapter 5.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

In educational and psychological measurement, often the primary goal is to measure a latent variable such as reading proficiency or anxiety. Item Response Theory (IRT), which is the dominant paradigm in measurement today, offers a family of mathematical models to specify the relationship between the latent trait of interest and the test items that are designed to measure this trait. In educational measurement, the position of an item on the latent dimension of interest is called item difficulty. Each examinee is also located on the same latent dimension and their position is referred to as their proficiency. In its simplest form, the IRT model gives the probability of a correct response to an item in terms of the item difficulty and examinee ability. In its more general forms, unidimensional IRT models address item discrimination, examinee guessing, and polytomously scored items.

While IRT is principally concerned with the relationship between item characteristics, latent examinee traits, and observed responses, it has been shown that covariate data also has a potential role in test validation, test development, and parameter estimation (e.g., Justice, Bowles, & Skibbe (2006); Smith (2000); Chang, Plake, & Ferdous (2005)). Model parameter covariates (sometimes called collateral information) may help measurement specialists understand *why* they observe the values of the parameters in their models. In other words, covariates offer researchers an opportunity to investigate some of the potential reasons for observed differences across items and examinees. For example, given person covariates such as subgroup membership, gender,

or education, the extent that the ability parameter is related to these covariates can be estimated, which may provide valuable consequential validity evidence. Similarly, item or testlet covariates such as word count or presence of graphic representation could help explain why an item is harder or more discriminating compared to others. Such information could help inform item writing, item selection, or even item parameter estimation. Additionally, covariates may reduce unexplained variance and result in greater precision in estimation. Two familiar examples that illustrate how covariates are used in augmenting estimation are the U.S. National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS). In these tests, examinees are administered relatively short tests in various subject areas such as mathematics, along with numerous background questions (covariates). The covariates then contribute to the estimation of the posterior distributions for examinee ability. This strategy is employed to improve measurement precision over using test performance alone.

The potential value of covariate data has been shown; of interest here is in *estimating* the covariate relationships. The dominant approach to doing this uses post-hoc regression analyses. There are, however, shortcomings to this approach, which will be discussed below, that have led Wainer, Bradlow, and Wang to propose incorporating covariates *directly* into the measurement model (2007). This new framework, the fully Bayesian Testlet Response Theory (TRT) model with covariates, simultaneously estimates IRT parameters and their covariate coefficients. The potential advantages of this approach will be presented in the next section. First, while separate from the issue of covariates, the impetus for this model, the testlet, should be introduced.

IRT models typically rely on an important mathematical assumption: *local independence*; that is, after accounting for examinee proficiency, responses should be statistically independent across items. One frequent violation of local independence stems from testlet structures. A testlet is a set of items that utilize a common stimulus (passage, picture, graph, etc.). Because testlet items share the same stimulus, a dependency may arise across items. When this occurs, acknowledging the underlying testlet structure in a test is important for modeling the test data appropriately. Otherwise, ignoring the testlet structure may result in underestimated standard errors for proficiency parameters—and thus overestimation of precision—and biased item parameters (Thissen, Steinberg, and Mooney, 1989; Wainer & Thissen, 1996; Sireci, Thissen, and Wainer, 1991; Keller, L. A., Swaminathan, H., and Sireci, S. G., 2003). One approach to model this dependency is to treat each testlet as a polytomous item. This approach of using polytomous models to handle local dependence has been widely researched and despite the value that the research has shown, it has not been widely adopted in practice, perhaps due to some of the limitations. First, when testlets are treated as polytomous items, a single discrimination parameter is estimated for all items within a given testlet instead of estimating a unique discrimination parameter for each item. This may result in loss of potentially valuable information. Another equally important limitation is that polytomous scoring ignores response *patterns* within the testlets, which again may contain very valuable information.

Another approach, which addresses the shortcomings of the polytomous scoring approach, is using a parametric framework called Testlet Response Theory (TRT; Bradlow, Wainer, & Wang, 1999). In TRT, an item nested within a testlet remains the

unit of measurement, which addresses both the common discrimination parameter and the pattern scoring issues that characterize the polytomous scoring approach. To model the dependency between the items, TRT incorporates a random effect parameter into the familiar IRT models. This parameter accounts for the shared variance among items within a testlet. Research has repeatedly shown its value using both simulated and real data (e.g., Bradlow, Wainer, & Wang, 1999; Wang, Bradlow, Wainer, 2002).

Wainer and his colleagues' fully Bayesian TRT model with covariates differs from standard IRT models by using a fully Bayesian hierarchical framework to model the data. There are several important advantages to using a fully Bayesian hierarchical framework over likelihood methods in this context. First, it permits the inclusion of prior knowledge about the test and the examinees in the model, while allowing sharing of information across persons, items, and testlets, which results in improved estimation of the parameters. Second, having the entire posterior distribution of the parameters of interest allows for making probabilistic inferences in a very simple and intuitive manner. Third, it allows covariates to be estimated simultaneously with the item and person parameters. Within this framework, one can incorporate all the prior information and/or beliefs regarding the parameters along with their covariates into a single measurement model.

1.2 Statement of Problem

Understanding the relationship between person, item, and testlet covariates and person, item, and testlet parameters may offer considerable benefits to both test development and test validation efforts. Indeed, the Standards for Educational and Psychological Testing (1999) recommends investigating validity evidence based on relationships with external variables noting that "[c]ategorical variables, including group membership variables, become relevant when the theory underlying a proposed test use suggests that group differences should be present or absent if a proposed test interpretation is to be supported" (p.13). The idea of investigating relevance of categorical membership could easily be extended to differential item and test functioning, which are also among top concerns of the researchers and practitioners in testing.

Covariate information may also help test developers understand what characteristic of items are related to item discrimination and difficulty. Having such information could potentially save time and resources during test development. For example, it is conceivable that such information may be useful in selecting or writing items for certain purposes or could even be used to generate item-specific priors to improve estimation under small sample conditions (e.g., Keller, 2002; Baldwin, Keller, Hambleton, 2004).

The Bayesian TRT models proposed by Wainer et al. offer a unified structure within which model parameters (including testlet parameters when necessary) may be estimated simultaneously with model parameter covariates. This unified approach represents an important advantage of these models: theoretically correct modeling of the relationship between covariates and their respective model parameters. Analogous

analyses could be performed via conventional post-hoc regression methods, however, the fully Bayesian framework offers an important advantage over the conventional post-hoc methods by reflecting the uncertainty of the model parameters when estimating their relationship to covariates. In other words, the results of post-hoc regression analyses that use point-estimates as dependent variables could be misleading because the regression model ignores that these values are *estimated* (or assumes that they can be treated as true). With the help of the Bayesian framework, the TRT models overcome this problem: covariates are directly included into the measurement model, which in turn allows researchers the opportunity to model the relationships between model parameters and covariates correctly by accounting for the uncertainty in the model parameters when estimating their relationship to covariates.

1.3 Purposes and Significance of Study

The purpose of this study was twofold. First, to conduct a basic simulation study to show that the Bayesian TRT approach is functioning as expected in estimating the relationship of covariates to their respective model parameters. Extending the hierarchical Bayesian TRT model to incorporate covariates is a relatively new development and has not been evaluated using simulated data. Thus, the primary focus of the simulation study was to evaluate the model's functioning. For this purpose, simulation conditions were arbitrarily limited to ability parameter covariates. Additionally, the Bayesian TRT results were compared to post-hoc regression results, where the dependent variable was the point estimate of the model parameter of interest which is the proficiency parameter in this research. For post-hoc analyses, new point estimates were obtained from a new set of

calibrations that did not include covariates in the model. This was necessary since the covariate coefficients are simultaneously estimated with the model parameters and therefore, would be a confounding factor in the subsequent regression analyses where the point estimates (the mean of parameters' respective posteriors) were estimated using the very same covariates.

Second, an empirical study applied the Bayesian TRT model to two real data sets: the Step 3 component of the United States Medical Licensing Examination (USMLE™), and the Posttraumatic Growth Inventory (PTGI) by Tedeschi and Calhoun (1996, see Appendix). The Step 3 exam is in typical multiple-choice format while the PTGI is a 21-item 5-point Likert-type survey designed to address changes people report following a traumatic event. In this instance, the survey was administered to women recently diagnosed with in breast cancer. In addition to examinee responses, background questions and item-specific collateral information were also collected for both assessments.

To evaluate both empirical studies, conventional post-hoc analyses were also conducted on the data sets, using the point estimates of the parameters as the dependent variables. In so doing, the extent to which the two strategies (simultaneous or post-hoc estimation) yielded different results could be evaluated. Thus, contrasting using covariates within a Bayesian framework with the more common post-analysis regressions was the main rationale of these studies. Finally, as both data sets contain testlets, the effect of not modeling this dependency on ability estimates was also investigated.

CHAPTER 2

LITERATURE REVIEW

This review contains three major sections. First, the basic principles of the Testlet Response Theory (TRT; Bradlow, Wainer, & Wang, 1999) are described in a stepwise fashion, starting with the tenets of the traditional Item Response Theory (IRT) and continuing with a review of the research on different strategies for modeling local item dependence in an IRT framework. The next section presents an overview of Bayesian and frequentist approaches to estimation and inference in the context of psychometrics. Relevant studies comparing the two approaches are also reviewed here. The final section discusses the value that covariate data may have for measurement scientists including applications in estimation, test development, and validation. Included is a comparison of the frequentist approach to modeling covariate relationships with the approach taken using a fully Bayesian TRT model.

2.1 Item Response Theories and the Issue of Local Dependence

Modern test theory is grounded largely in seminal work of Frederic M. Lord in Item Response Theory (1952). Later, Birnbaum (1968) put forward an item response model that was mathematically in the logistic form rather than Lord's normal ogive form. IRT defines the probability of success (or endorsement) on a given item as a function of a respondent's proficiency and the item's characteristics. The function for the *three-parameter logistic model* (3PL), which is the most general dichotomous model in widespread use, is given in equation 1 below:

$$P_i(u_i = 1 \mid \theta, a_i, b_i, c_i) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \quad (2.1)$$

where $P_i(u_i = 1 \mid \theta, a_i, b_i, c_i)$ is the probability of an examinee with a proficiency of θ having a response $u_i = 1$ to item i b_j is the item difficulty parameter, a_j is the item discrimination parameter, c_j is the pseudo-guessing parameter, e is the base of the natural logarithm equal to approximately 2.718, and D is a scaling constant equal to 1.7, which is used so that the logistic function approximates the normal ogive function.

The *two-parameter logistic model* (2PL) and *one-parameter logistic model* (1PL) may be viewed as special cases of the 3PL. For both the 2PL and the 1PL, the pseudo-guessing parameters (i.e. the *c*-parameters) are fixed to zero. In addition, for the 1PL the discrimination parameters (i.e. *a*-parameters) are constrained to be equal across all items (in some cases fixed to 1 by convention).

When response data are not binary, *polytomous* IRT models are used. Various polytomous models have been developed along similar but slightly divergent lines. Perhaps the most widely used is the Graded Response Model (Samejima, 1969), which is a generalized form of the 2PL wherein each item is defined by a single difficulty parameter and multiple threshold parameters—one for each score category (excepting the minimum score category). The graded response model describes the probability of obtaining a particular score or higher on a given item as a function of theta:

$$P_{i,x}{}^{*}(u_i \geq x \mid \theta, a_i, b_i, c_i) = \frac{e^{Da_i(\theta - b_{i,x})}}{1 + e^{Da_i(\theta - b_{i,x})}} \qquad (2.2)$$

where $P_{i,x}{}^{*}(u_i \geq x \mid \theta, a_i, b_i, c_i)$ is the probability of obtaining a score of x or higher b_{ix} is the location parameter, and a_i, e, and D have the same interpretation as in equation 1. Given equation 2, the probability of getting score x is given by:

$$P_{i,x}(u_i \geq x \mid \theta, a_i, b_i, c_i) = \frac{e^{Da_i(\theta - b_{i,x})}}{1 + e^{Da_i(\theta - b_{i,x})}} - \frac{e^{Da_i(\theta - b_{i,x+1})}}{1 + e^{Da_i(\theta - b_{i,x+1})}}. \qquad (2.3)$$

That is, the probability of scoring in a specific score category is equal to the probability of scoring in that category or higher minus the probability of scoring higher.

The IRT models presented above rely on two important assumptions. First, they assume *unidimensionality*; that is, they assume that only one dominant trait is being measured by the items of a given test. Second, as mentioned above, they assume *local independence* or that after controlling for proficiency, "examinees' responses to any pair of items are statistically independent" (p.10, Hambleton, Swaminathan, & Rogers, 1991). Hambleton et al. explain that when unidimensionality assumption holds, local independence is observed; however, local independence can be observed when the data is multidimensional. That is, when all ability dimensions are accounted for, one would observe local independence. Local independence does not hold if this is not the case.

One of the most common situations where local independence is in jeopardy is when tests are composed of testlets. Wainer & Kiely (1987) define a testlet as "a group of items related to a single content area that is developed as a unit and contains a fixed

number of predetermined paths that an examinee may follow" (p.190). It is well established in the literature that ignoring the testlet structure in the test may violate the local independence assumption and result in underestimated standard errors of ability parameters and biased item parameters. For example, using a reading comprehension exam with four passages and 22 associated questions (with 7, 4, 3, and 8 items), Thissen, Steinberg, and Mooney (1989) calibrated the data with Bock's (1972) nominal model and compared it to the conventional 3PL analysis. They showed that modeling the testlet structure as polytomous responses results in *less* precision than the conventional IRT analysis estimated, and interestingly, similar or higher concurrent validity was observed when the testlet-based scores were correlated with an external criterion. They explained that the 3PL estimation of precision was essentially wrong, because it incorrectly assumes local independence and therefore overestimates the information.

In a related study, Sireci, Thissen, and Wainer (1991) showed the detrimental impact of unaddressed local dependence on reliability. The authors computed the marginal reliability for 45 SAT pretest items composed of 4 testlets (12, 13, 10, and 10 items) using conventional 3PL and Bock's (1972) nominal model. They found that, with the 3PL estimation reliability was overestimated by 10-15%. Using the Spearman-Brown formula, they showed that the test length needed to be doubled (8 testlets) to close the 15% gap.

Various strategies have been proposed to handle testlet structures. Unfortunately, a commonly used approach is to ignore the local dependency problem and use traditional IRT models –perhaps inappropriately. As the research by Thissen et al. (1989) and Sireci et al. (1991) reviewed above suggested, failure to model local dependency may create

positive bias in test information, which may be especially problematic in the context of adaptive testing. Thus, when local dependence exists, it may be desirable to use a measurement model that takes the dependency between items into account. Thissen et al. (1989) utilized a polytomous IRT model that was proposed by Bock (1972) to handle local dependence. Here, the polytomous model defines the testlet as the unit measure measurement, treating the testlets as conditionally independent items rather than modeling conditionally dependent items. Such an approach creates a polytomous item with a score ranging from zero to the total number of items associated with the stimulus, eliminating the dependency problem.

The approach of using polytomous models to handle local dependence has been widely researched and proven valuable. For example, Wainer (1995) examined local dependency and its impact on score precision for the Reading Comprehension and Analytic Reasoning sections of the Law School Admission Test. He found that the dependency in the testlet structure had an important effect on the statistical characteristics of the test such as reliability and that this effect was not captured when the data were modeled incorrectly by assuming independence. When the dependency was modeled, the reliabilities of these sections were found to be considerably lower. In a more recent study, Zenisky, Hambleton, and Sireci (2002) compared the reliability estimates obtained using a dichotomous and polytomous scoring models using Medical Collage Admission Test data and observed substantial differences. The Spearman-Brown calculations showed that the Verbal Reasoning section of the test needed to be 50% longer to achieve the reliability calculated by the dichotomous approach.

When testlets are present, utilizing a measurement model that takes the dependency into account seems warranted in many cases. However, Zenisky et al. (2002) point out that there are (at least) two potential disadvantages when the polytomous IRT approach is taken. First, such models require a common discrimination parameter for all items within a testlet, which may result in loss of potentially valuable information. Second, this approach relies on the sum score within each testlet, which ignores any differences between scoring *patterns*. Thus two examinees with the same number-correct score for a given testlet are regarded alike despite the fact that they may have answered different questions correctly. Hence, the decision between dichotomous and polytomous approaches may be difficult even when violations of local dependence is suspected.

An alternative approach is using a parametric framework called Testlet Response Theory (TRT; Bradlow, Wainer, & Wang, 1999). In TRT, an item nested within a testlet remains the unit of measurement. To accomplish this, TRT incorporates a random effect parameter (γ) to the familiar models presented above in equations 1 and 2. This parameter accounts for the shared variance among items within a testlet and it does so via a *fully* Bayesian hierarchical framework (i.e., priors are specified for all parameters in the model). The next section describes TRT models and reviews the literature that has used them.

2.2 Testlet Response Theory Models

TRT provides a psychometric model for calibrating and scoring tests composed of testlets within a fully Bayesian framework (Bradlow, Wainer, & Wang, 1999). The 3PL TRT model is a simple extension of the standard 3PL where an additional parameter γ is

incorporated to model the local dependence. The function for the *three-parameter TRT model* is given in equation 4 below:

$$P_{ij}(\theta_i) = c_j + (1 - c_j)\frac{e^{Da_j(\theta_i - b_j - \gamma_{id(j)})}}{1 + e^{Da_j(\theta_i - b_j - \gamma_{id(j)})}} \tag{2.4}$$

where $\gamma_{id(j)}$ is the testlet effect for person i and item j nested within testlet *d*. Within this framework, Samejima's Graded Response Model (1969) becomes:

$$P_{ij,x}{}^{*}(\theta_i) = \frac{e^{Da_j(\theta_i - b_{j,x} - \gamma_{id(j)})}}{1 + e^{Da_j(\theta_i - b_{j,x} - \gamma_{id(j)})}}, \tag{2.5}$$

where $P_{ij,x}{}^{*}(\theta_i)$ is the probability of examinee i getting score x or higher on item j nested within testlet *d*. The probability of obtaining a specific score of x is given by equation 6:

$$P_{ij,x}(\theta_i) = \frac{e^{Da_j(\theta_i - b_{j,x} - \gamma_{id(j)})}}{1 + e^{Da_j(\theta_i - b_{j,x} - \gamma_{id(j)})}} - \frac{e^{Da_j(\theta_i - b_{j,x+1} - \gamma_{id(j)})}}{1 + e^{Da_j(\theta_i - b_{j,x+1} - \gamma_{id(j)})}}. \tag{2.6}$$

Bradlow et al. (1999) introduced the Bayesian TRT model, its computation, and demonstrated its accuracy and effectiveness via a 2 x 3 factorial simulation study. The number of examinees (I=1,000), test length (J=60), and percentage of items nested within testlets (50%) were held constant across study conditions. Model-based data were

generated using the two-parameter TRT model. The simulating item parameter distributions were chosen to match the SAT marginal distributions to make results realistic and relatively comparable to analyses using real SAT data that were discussed in the paper. The effect of two factors, number of items within each testlet (5 or 10) and testlet variance (σ_γ^2=0.5, 1, 2) were tested. Additionally, a control condition was simulated in which all items were independent (i.e., σ_γ^2=0) to serve as a point of for comparison and to demonstrate that the model can detect the absence of testlet effect and find the same solution as the traditional approaches. The six study conditions and the control condition were analyzed in three ways: via BILOG (Mislevy & Bock, 1983) with the assumption of conditional independence, a Data Augmented Gibbs Sampler (DAGS) approach without the testlet effect parameter (γ), and finally a DAGS approach with γ (DAGS γ). The results of the simulation study confirmed the accuracy and effectiveness of the DAGS γ approach. In fact, for all parameters, the mean absolute error for DAGS γ was less than for DAGS and BILOG across all conditions including the control condition. The DAGS and BILOG approaches performed very similarly across all conditions and all three approaches had increasingly higher mean absolute error as σ_γ^2 increased. The rank correlations between the estimates and true values for the ability parameter were also computed. Not surprisingly, DAGS γ estimates' correlations were higher than the other two approaches for ability and discrimination parameters. Correlations for the difficulty parameter were approximately the same for all methods across conditions. Lastly, the Mean 95% Posterior Interval Width values revealed that DAGS posterior intervals for the ability and discrimination parameters were narrower

than DAGS γ across all conditions and the discrepancy between the two methods increased as the amount of σ_γ^2 increased. This finding is also expected given the overestimated posterior information when the testlet effects are not taken into account.

In a follow up study, Wang, Bradlow, and Wainer (2002) extended the dichotomous TRT approach to mixed-format tests. The study was composed of a simulation study and two applications using operational data from the Test of Spoken English and the North Carolina Test of Computer Skills. The simulation component of the study examined the success of the model in recovering the true parameters. Three factors were manipulated: Number of categories for each item (2, 5, 10), testlet length (3, 6, 9), and testlet variance (σ_γ^2=0, 0.5, 1). Response data for 1,000 simulees were simulated for a 30-item test across five replications for each condition. Of the 30 items, 12 were independent dichotomous items, and 18 were testlet items, either dichotomous or polytomous. The total number of testlet items was fixed to 18. Again, model parameters were drawn from SAT's parameter distributions to obtain realistic values. Finally, to sample from the posterior distributions, the data augmentation approach (Tanner & Wong, 1987; as cited in Wang et al., 2002) was used for item thresholds and the Metropolis-Hasting method (Hasting, 1970) was used for the other model parameters.

Wang et al.'s (2002) findings were parallel to Bradlow et al.'s (1999). Authors used two criteria to evaluate the success of the model: correlation between the true and the estimated parameters and the Mean Square Error (MSE) of the estimates from the true values. The results of both evaluation criteria indicated that the model recovered the difficulty and threshold parameters very well (average of r= 0.99 and 0.98, respectively). The average correlations for the ability, discrimination, and guessing parameters were

0.93, 0.89, and 0.60, respectively. Even though these correlations are not as encouraging as the difficulty and threshold correlations, the authors point out that they are in line with the findings of the previous simulation studies in the literature. Next, Wang et al. applied this approach to the operational data sets mentioned above. The results of the North Carolina Test of Computer Skills showed considerable testlet effect for one section, while less significant testlet effects were observed for others. When the authors examined the plot of the three information functions obtained via the three different approaches, the result was striking. When all items were considered conditionally independent, as they often are, the peak of the information curve was much higher than the other methods, indicating overestimation of information. When the testlet effect was taken into account via the TRT model, the peak of the curve was lower and the curve itself was much flatter. Finally, the last curve was obtained by converting each testlet into a polytomously scored item. This method resulted in an *underestimation* of information—50% for most of the theta scale. Lastly, the authors fitted the TRT model to the Test of Spoken English. The findings indicated that, at least for the data used in the study, the assumption of conditional independence was essentially met. Wang et al. concluded that the current practice of calibrating the Test of Spoken English with independence assumption was acceptable.

The results of these two studies indicate that when testlet effects are present, the TRT approach models this effect better than other models in use, resulting in improved estimation of parameters and their standard errors. Another benefit of these researchers' work with the TRT model is not due to the model itself but rather their approach to estimating it. Next, details about the fully Bayesian estimation of the TRT model will be

provided, including its general principles, its merits, and its differences from the traditional frequentist estimation methods.

2.3 Bayesian vs. frequentist Methods

There are two main schools of thought in inferential statistics: frequentists and Bayesians. In the frequentist paradigm, population parameters are viewed and treated as unknown but *fixed* quantities. The ultimate purpose is to set confidence intervals around the point estimates of the parameters of interest and conduct hypothesis testing via an appropriate null hypothesis. In the Bayesian paradigm, the population parameters are considered unknown and *random*. In the Bayesian framework, the goal is to obtain posterior distributions for the parameters of interest through the likelihood and prior distributions and to conduct hypothesis testing by constructing Bayesian confidence intervals and posterior *p*-values. Little (2006) describes these two approaches as:

> "…I regard a "frequentist" as one who bases inference for an unknown parameter θ on hypothesis tests or confidence intervals, derived from distribution of statistics in repeated sampling. I regard a "Bayesian" as one who bases inferences about θ on its posterior distribution, under some model for the data and prior distribution for unknown parameters" (p.214).

Thirty years ago, debates regarding which paradigm provided the "right" way of doing statistics were both philosophical and methodological. Today, with the advances in computing speed and developments in the Monte Carlo techniques, Bayesian methods

have proved themselves to be scientifically sound and quite useful for many applications. Nevertheless, growing recognition of the utility of Bayesian methods does not mean these philosophical differences have been resolved. Both paradigms have been criticized for various reasons: for instance, frequentist methods for being atheoretical, and Bayesian methods for being subjective and requiring researchers to provide greater model specification (Little, 2006).

In their recent book, Wainer, Bradlow, and Wang (2007) address the criticism regarding the subjectivity of the Bayesian paradigm. Specifically, most frequentists are concerned that when an improper prior is chosen it may bias the posterior distribution and hence the inferences being made. Wainer et al. point out that placing priors on the likelihood parameters is "fundamentally correct" given that the likelihood parameters are unknown and random. They add that via the priors, one could incorporate additional information (obtained from former research studies, other data, etc.) into the model. Furthermore, they point out that it is up to the researcher to determine how informative the prior is going to be; for example, one could choose to put uninformative priors that would yield posterior distributions that are proportional to the likelihood. In any event, as Wainer et al. argue, the likelihood could also be wrongly specified and needs model checking just as the priors do.

2.4 Bayesian Inference and Computation in IRT

The most popular and widely used frequentist approach to IRT parameter estimation is the marginal maximum likelihood (MML) estimation (Bock & Lieberman, 1970). MML parameter estimation is performed in two stages. In the first stage, the

ability estimates are considered as a random sample from a population distribution and are integrated out over this distribution to maximize the marginal likelihood for item parameters (Hambleton, Swaminathan, & Rogers, 1991). As Hambleton et al. note, the resulting item parameters have asymptotic properties; that is, estimates are consistent as the examinee number increases. In the next step, the item parameters are treated as known and the ability parameters are estimated. Similarly, more items yield better ability estimation. Specifically, given an *N*-item response vector U for examinee j, the likelihood function for the 3PL may be written as:

$$P(U \mid \theta, a, b, c) = \prod_{i=1}^{N} P_i^{U_i} Q_i^{1-U_i} . \qquad (2.7)$$

It follows that,

$$P(U, \theta \mid a, b, c) = \prod_{i=1}^{N} P_i^{U_i} Q_i^{1-U_i} g(\theta) , \qquad (2.8)$$

and the marginal probability of obtaining response pattern U (π_U) can be expressed as:

$$\pi_U = P(U, \theta \mid a, b, c) = \int_{-\infty}^{\infty} \prod_{i=1}^{N} P_i^{U_i} Q_i^{1-U_i} g(\theta)\, d(\theta) . \qquad (2.9)$$

(Hambleton & Swaminathan, 1985, p.140)

where, $g(\theta)$ is the assumed population distribution of θ_i (e.g., $\theta_i \sim N(0,1)$).

There are, however, known limitations of MML estimation. An important one is the challenge of estimating the discrimination and guessing parameters in the 3PL. For the 3PL, more than one solution may approximately fit the model and the data may not be able to differentiate among them (Wainer et al., 2007). In some cases, this may result in unreasonable estimates (e.g., estimates drifting out of bounds). Another important limitation is MML estimation's reliance on asymptotic theory, which leads to difficulties

when sample sizes are small. The effectiveness of Bayesian solutions to such problems has been documented by many studies in the literature. Some of these studies are reviewed next.

Swaminathan and Gifford (1982, 1985) developed conditional and joint hierarchical Bayesian IRT models and demonstrated that Bayesian estimates were indeed superior to the maximum likelihood estimates. In their 1982 study, Swaminathan and Gifford introduced a Bayesian method in which parameters are estimated either by joint maximization with respect to the parameters of the posterior density or by conditioning on the difficulty. To evaluate this method, they conducted a two-part simulation study. In the first part, they compared ability estimates for the 1PL obtained via the conditional Bayesian approach to those obtained using the conditional maximum likelihood approach. Effects of two factors were examined: sample size (20, 50) and number of items (15, 25, 40, 50). They reported that performances of both methods were highly similar in terms of the correlation between the true theta values and their respective estimates, except when sample sizes were small or test lengths were short. For small samples or short tests, Bayesian method produced better correlations. They also found that the mean squared deviations for the Bayesian estimates were considerably smaller than the maximum likelihood estimates' and the discrepancy was particularly great for small sample sizes and short tests.

In the second part of the study, they compared the joint maximum likelihood and joint Bayesian estimates of the ability and difficulty parameters for the 1PL. Bayesian estimation uses prior information about the distribution of the model parameters to improve estimates. The effect of the prior distribution on the estimates was also

examined. The results showed that the correlations for the difficulty parameter estimates were identical for the two methods; however, for ability estimates, the Bayesian method yielded somewhat higher correlations with the true values. The mean squared deviations for theta were again smaller for the Bayesian method. The authors noted that for large sample sizes, the difference between maximum likelihood and Bayesian estimates were negligible. For smaller sample sizes, however, the Bayesian method produced better results across the board. Lastly, their findings indicated that varying priors did not have an effect on the correlations with the true values; however, the mean squared deviations accuracy was affected, particularly in small samples and more so for ability estimates than item difficulty estimates. Swaminathan and Gifford concluded that Bayesian procedure was more accurate that the maximum likelihood methods and its use yielded more meaningful results especially for under extreme conditions (e.g., perfect scores, or zero correct responses).

Subsequently, Swaminathan and Gifford (1985) extended their approach with the 1PL to the 2PL. In a simulation study, effects of four examinee sample sizes (50, 100, 200, 500) and three test lengths (15, 25, 35) on Bayesian and maximum likelihood estimates were tested. The priors for ability and difficulty parameters were uniform to show the efficacy of the Bayesian estimation. A slightly informative prior was placed on the discrimination parameters. Their findings once again showed the superiority of the Bayesian procedure: in general, correlations with the true parameters were higher and the mean square differences were lower. They note that as the number of items and the number of examinees increase, the two methods start yielding very similar results.

In a recent simulation study, Gao and Chen (2005) compared the MML approach to

Bayes modal estimation in the context of the 3PL. The study also examined the effect of different priors on the Bayesian estimates and the robustness of the Bayesian procedure to discrepancies between the prior means and the true parameters. Their results were in line with Swaminathan and Gifford's findings (1982, 1985). They observed that the Bayesian estimates were generally more accurate, resulting in higher correlations and lower RMSDs with the true values. Additionally, an interaction between the impact of prior specification and sample size was observed. For small sample sizes, the impact of the prior on discrimination and guessing parameters were, at times, considerable. However, this effect diminished as the sample size increased. Lastly, when the sample size was large, the likelihood and Bayes approaches produced very similar results.

Given its performance relative to maximum likelihood procedures, it is no surprise that the Bayesian approaches for estimation are becoming increasingly popular. Gelman et al. (1995) nicely summarize Bayesian data analysis in three steps: (i) setting up a full probability model (a joint distribution for all parameters in the model observable and unobservable), (ii) conditioning on observed data (calculating the posterior distribution which is conditioned on the observed data), and (iii) evaluating the fit of the model.

Hierarchical Bayesian models are utilized when information on different observational units is available. For hierarchical Bayesian models, Wainer, Bradlow, and Wang (2007) partition model specification step into three parts: the likelihood, the prior and the hyperprior. The likelihood function gives the likelihood of parameter values conditioned on the observed data. Using preformed beliefs or information, prior distributions are formulated that provide information about the likelihood parameters. This information could be obtained from other relevant research or may be based on the

researcher's belief about how they should be distributed. And lastly, to complete the hierarchical model, hyperpriors are placed on the priors to reflect their uncertainty. Both priors and hyperpriors are usually chosen to ensure a *proper* posterior distribution (i.e., posteriors that integrate to one). It is also desirable for priors and hyperpriors to be *conjugate* for the posterior (i.e., they follow the same parametric form) so that the posterior is in a tractable form and the inferences can be performed in a straightforward manner. However, in most problems, conjugate priors and hyperpriors do not exist and additional strategies must be employed to sample from the posteriors—e.g., Markov chain Monte Carlo (MCMC) techniques.

One of the most popular MCMC algorithms is the Metropolis-Hastings algorithm. Gelman et al. (1995) summarize the Metropolis-Hastings algorithm in 2 general steps:

1. Draw a starting point θ^0 from a starting distribution $p_0(\theta)$
2. For time t=1, 2, 3, ….T:
    a. Sample θ^* from a jumping distribution J at time t
    b. Calculate the ratio of densities $r = \dfrac{p(\theta^* \mid y)\, J_t(\theta^* \mid \theta^{t-1})}{p(\theta^{t-1} \mid y)\, J_t(\theta^{t-1} \mid \theta^*)}$
    c. Set $\theta^t = \theta^*$ with the probability of min(r, 1), and otherwise keep old value.
    d. Repeat (steps *a* through *d*) for T cycles

In summary, each iteration, the algorithm generates a random point (dependent on the previous draw only) whose stationary (target posterior) distribution is $P(\theta \mid y)$ and chooses the new value over the old value if it is more likely given the observed data.

Gelman et al. (1995) note that iterative simulation procedures like MCMC may create three complications. First, if too few iterations occur (that is, if T is not large

enough), the target distribution may not be reached. To examine whether the target distribution is reached, convergence must be tested. This is usually accomplished by examining the variation within and between simulated sequences (chains) until the within is approximately equal to between variance. This statistic is referred as $\sqrt{\hat{R}}$ and convergence is achieved when $\sqrt{\hat{R}} \approx 1$ (Gelman & Rubin, 1993). Second, even if the target distribution is reached by draw *T*, earlier draws will be sampled from pre-target distributions and thus the total sample of draws will not represent the target distribution. Gelman et al. recommend discarding some of the initial observations to moderate the impact of the starting distribution. This is usually referred as *burn-in period* in the Bayesian literature and the number of samples to be discarded varies from problem to problem. Third, because the draws are serially correlated, to have approximately independent draws from the target distribution, Gelman et al. suggest using only every k^{th} draw after convergence is achieved.

The efficiency and accuracy of the MCMC approach to Bayesian estimation has been demonstrated by many studies in the literature. Patz and Junker (1999a) applied an MCMC method that used a *Metropolis-Hasting within Gibbs* algorithm to fit the 2PL. This algorithm samples from complete conditional distribution according to *Gibbs* algorithm, but for conditional distributions whose forms are known up to a normalizing constant, uses an iteration of *Metropolis-Hasting* algorithm (Patz & Junker, 1999a). Using the National Assessment of Educational Progress (NAEP) Grade 4 reading data, they illustrated how well their approach performed compared with BILOG (Mislevy & Bock, 1985). In a subsequent study, Patz and Junker (1999b) successfully extended their approach to mixed-format tests including, 2PL, 3PL, and Master's partial credit model

and as well as addressing the missing response and multiple raters problems.

Kim (2001) compared an MCMC approach using Gibbs sampling to joint-, conditional-, and marginal maximum likelihood procedures (both expected a posteriori (EAP) and maximum likelihood were used under MML method to estimate theta) in the context of 1PL. Four data sets were used: (1) the Law School Admission Test (LSAT) data with 5 items and 1,000 examinees; (2) memory test data containing with 10 items and 40 examinees; (3) the 1992 NAEP Grade 4 reading test with 6 short constructed-response questions and 3,000 examinees; and (4) the English Usage Test with 31 items and 365 examinees. Kim reported that the difficulty parameter estimates were very similar across estimation methods for all four datasets. In fact, for the LSAT data and the English Usage Test, the difficulty estimates were essentially perfectly correlated across methods. The ability parameter estimates obtained by the Gibbs sampling and maximum likelihood/EAP were also very similar for all datasets (both obtained via a normal prior). Conditional and joint maximum likelihood approaches yielded very similar estimates but that were different from Gibbs and EAP estimates.

In a recent simulation study, Wollack, Bolt, Cohen, and Lee (2002) compared the accuracy of item parameter estimates for MML as implemented by MULTILOG (Thissen, 1991) and MCMC with Gibbs sampling as implemented by the BUGS computer program (Spiegelhalter, Thomas, Best, & Gilks, 1997) in the context of the nominal response model (Bock, 1972). Two factors were varied: number of items (10, 20, and 30) and number of examinees (300 and 500). Wollack et al. found that the item parameter recovery for both approaches were extremely similar. Both methods produced good estimates, even for fairly short tests and small sample sizes. Parameter recovery was

best for items with moderate difficulty and parameter recovery improved as test length increased from 10 to 30 items. The authors concluded that the MCMC method is a good substitute for MML estimation, particularly when MML algorithms are not available.

The research reviewed above supports the Bradlow et al.'s (1999) motive for taking a Bayesian approach. Recently, Wainer, Bradlow, Wang (2007) extended the Bayesian TRT model by incorporating item, person, and testlet covariates into the model to help understand/explain *why* we observe the values of the parameters in our models. The next section presents the rationale for modeling these data in a fully Bayesian framework.

2.5 Bayesian Testlet Model with Covariates

The TRT model proposed by Wainer, Bradlow, Wang (2007) differs from standard IRT models in three important ways: it incorporates a random effect parameter to the standard IRT model to account for the shared variance among items within a testlet (discussed above), it utilizes a fully Bayesian hierarchical framework (also discussed above), and it incorporates covariates directly into the model. In general, covariates may help illuminate relationships between model parameters and collateral information such as examine group membership or item readability. These relationships help researchers answer questions of *why* (Wainer et al., 2007). However, incorporating covariates directly into the measurement model offers additional benefits. Firstly, having covariates may improve parameter estimates. Second, it allows the relationships between model parameters and covariates to be modeled *correctly*. Understanding the variables that are associated with item, person, and testlet parameters is critical to both test development

and validation for obvious reasons. Traditionally, analyses are performed via conventional post-hoc regression methods; that is, after obtaining the point estimate for, say, theta one could regress it on the covariates and examine the relationships. For example, response time in computer-based testing is a variable whose relationship to model parameters has been heavily researched, and more often than not, these relationships are modeled via regression methods. A good example is Smith's (2000) study where he examined the relationship between item level response time and discrimination, difficulty, and word count, using data from the Graduate Management Admission Test. He performed a pairwise curvilinear regression analyses to model these relationships. Another good example can be seen in Chang, Plake, and Ferdous (2005). Using MANOVA, these researchers examined whether response times varied as a function of gender and US citizenship after controlling for proficiency in an adaptive test.

The post-hoc regression approach is easy to implement; however, it makes an untenable assumption. As Wainer et al. point out, conclusions drawn from post-hoc regression analyses using point-estimates as a dependent variable could be misleading because the model ignores that they are *estimates*, and hence, there will always remain a concern over bias. As they write:

> "…such conclusions are often unwarranted and careful statisticians move much more judiciously, deleting one variable at a time and examining how the pattern of regression coefficients changes, taking into account the inter-correlations among the independent variables and not deleting independent variables injudiciously" (p.178, Wainer et al., 2007).

So, it may be possible to get to the right answer via the traditional methods, however, compared to the Bayesian approach, it could be much more challenging. The Bayesian framework offers researchers the option of modeling all the parameters and their uncertainty simultaneously, sharing of information across items and people in a way that improves precision, and gives a more accurate picture of the relationship between the variables. The next section introduces and describes the model specification for the Bayesian TRT with covariates.

2.5.1 Model Specification

To fully specify a Bayesian model for the three-parameter TRT model, Wainer, Bradlow & Wang, 2007 specified the following distributions for the parameters of item *j* and person *i*:

For the item parameter vector $\zeta_j = \left(\log(a_j), b_j, \text{logit}(c_j)\right)$

$$\zeta_j \sim MVN\left(\mu_j, \Sigma\right)$$

$$\theta_i \sim N\left(W_i\lambda, 1\right),$$

$$\gamma_{id(j)} \sim N\left(0, \sigma^2{}_{d(j)}\right)$$

Covariates are incorporated into the model via the mean of the prior distribution of the item and ability parameters.

$$\mu_j = X_j^{'}\beta.$$

As it is evident from the above formulae, the βs and λ are the covariate slopes for the item and person parameters, respectively. If covariates for $\theta(W_i)$ exist, W_i will not have an intercept and be centered at zero to identify the model.

To insure proper posteriors, a set of slightly informative conjugate hyperpriors are also placed on the priors, completing the hierarchical Bayesian model: normal hyperprior on the vector of means and covariate slopes, an inverse-gamma hyperprior on the testlet variances, and an inverse-Wishart hyperprior on Σ, respectively, to ensure proper posteriors (Wainer et al., 2007).

2.6 Summary

It is not surprising that the Bayesian approach has become increasing popular in the past two decades. Computing power was the biggest hurdle the Bayesian methods faced and advances in computing power have made this problem more manageable. The advantages of using a fully Bayesian framework over the likelihood methods are considerable. To begin, Bayesian methods provide means to incorporate our knowledge/beliefs about the parameter distributions into the model. Wainer et al. (2007) note "There is no easier way of formally including everything we know into the scientific mechanism from which we can draw inferences" (p.113). Naturally, choice of priors has an impact on the accuracy of estimation, particularly for small samples (e.g., Swaminathan & Gifford, 1986). However, for the prudent researcher, it is easy to avoid making poor choices and take advantage of this powerful mechanism.

An equally important advantage is the capability of Bayesian procedures, with the help of MCMC methods, to allow the inclusion of the uncertainty in the item

parameters into ability estimation. The likelihood approaches do not account for the uncertainty of item parameter estimates in estimating ability. This is an issue for more complex psychometric problems such as those arising from small samples. Bayesian methods provide a straightforward solution—they do not rely on the asymptotic theory and they estimate parameters simultaneously.

Another key advantage of Bayesian methods is that it makes it possible to estimate parameters for examinees with unusual response patters (e.g., perfect scores). Similarly, in Bayesian framework, parameters for items with perfect response rate (or no correct response) can still be estimated. Bayesian estimation can also address the problem of item parameters estimates drifting out of range. And lastly, having the posterior distribution allows for making simple and intuitive probabilistic inferences about the model parameters. As Wainer et al. (2007) put "MCMC methods essentially turn inference into simply adding, counting, and sorting" (p.124).

In addition to the advantages listed above, the Bayesian TRT model presented above incorporates covariates into the measurement model in a very natural way: via the means of the prior distributions. Importance of covariates and their treatment is also discussed above. Given its technical and theoretical advantages, it is obvious why Bayesian framework was preferred for the TRT models with covariates (Bradlow, Wainer, & Wang, 1999; Wang, Bradlow, & Wainer, 2002; Wainer, Bradlow, Wang, 2007).

CHAPTER 3

METHODOLOGY

This study comprises two parts: a simulated data component and an empirical data component. The purpose of the simulation was twofold: First, the accuracy and effectiveness of two approaches to estimating covariate relationships—the Bayesian TRT approach and the more common post-hoc regression approach—were compared. Second, these findings helped inform the interpretation of the empirical results, since truth is not knowable using empirical data. The purpose of the empirical part of this study was to investigate the impact that the differences between these two approaches had on the inferences made about the covariate relationships.

This chapter is divided into three main sections: data description, parameter estimation, and data analysis.

3.1 Description of Data

In this section, study data are described in detail. Data from three tests were used: a simulated dataset, USMLE Step 3 dataset, and the PTGI data set.

3.1.1 Simulation Study Data

The main purpose of the simulation study was to compare the accuracy and effectiveness of the Bayesian TRT approach in estimating the relationship of covariates to their respective model parameters with the conventional post-hoc regression approach. As such, the only variable of interest was the covariates' relationship to the proficiency

parameter. The generating parameters were created using the built-in *rmvnorm* function of *S*-plus software package (MathSoft, Inc, 1999). This function returns a random sample of values from the multivariate normal distribution with a specified correlation matrix and mean and standard deviation vectors. The correlation matrix used in generating the data set mimics the relationship between the proficiency parameter and its covariates in a large-scale operational test and presented below.

$$corr = \begin{bmatrix} 1.00 & 0.67 & 0.38 & -0.70 & -0.20 \\ 0.67 & 1.00 & 0.31 & -0.36 & -0.32 \\ 0.38 & 0.31 & 1.00 & -0.22 & -0.32 \\ -0.70 & -0.36 & -0.22 & 1.00 & 0.30 \\ -0.20 & -0.32 & -0.32 & 0.30 & 1.00 \end{bmatrix}$$

The mean and standard deviations of each variable were set to 0 and 1, respectively. Using the *rmvnorm* function, a set of five random variables was simulated for 2,000 simulees. Next, the data set was divided into two subsets; the first one had the covariates that were highly correlated (0.67 and -0.70) with the proficiency parameter and the second data set had the covariates that were correlated very modestly with the ability parameter (0.38 and -0.20). The correlations between the two covariates within each data set were chosen to be approximately the same (-0.36 and -0.32). The two levels of covariate correlation—highly and modestly correlated—comprise the conditions of the simulation part of this study. The strength of correlation was varied to investigate the Bayesian TRT model's sensitivity in capturing the relationship between the variables. Once the data was generated, the resulting regression coefficients from the correlation matrix obtained from the data were calculated and are shown in Table 3.1.

The generating item parameters were obtained from a large-scale high-stakes high school mathematics exam. The test was composed of 50 items: 34 3PL items, 8 2PL items, and 8 polytomous items with 4 score categories each.

Once the generating item and proficiency parameters were obtained, the probability of a simulee answering an item correctly was calculated and compared to a random number sampled from a uniform distribution $U(0, 1)$. For dichotomous items, when the calculated probability for a given examinee was greater than the random number, then the simulee was assigned a 1 (correct response), if not, they were assigned a 0 (incorrect response). For polytomous items, Samejima's GRM (1969) was used. The polytomous decision rule in generating item responses was similar to the dichotomous case, except the random number was compared to cumulative probabilities of the four ordered categories. 50 replications were conducted for each study condition. 50 other replications without the covariates were also run to gather theta estimates without the covariates in the Bayesian model for use in the post-hoc regression analyses, using the data.

3.1.2 USMLE Step 3 Data

The data set used a sample of 112 multiple-choice items from Step 3 of the United States Medical Licensure Examination (USMLE®). Step 3 is designed to assess whether a physician possesses the qualities deemed essential to assume responsibility for providing unsupervised general medical care. This examination consists of two parts: 480 traditional multiple-choice items and nine performance assessment tasks developed to evaluate physicians' patient management skills through computer case simulations. For

the purposes of the present study, only dichotomous items were included in the analyses. To reduce computation time, 112 items were randomly sampled from a test form. The data sets contained 729 examinees. Of the 112 items, 50 were independent and the remaining 62 were nested within 28 testlets—22 of the 28 testlets had two items and the remaining six had three items associated with each.

Three item parameter covariates were also obtained for the analyses: vignette word count, stem word count, and options word count. The same three covariates were used for testlets, measured at the testlet level (sum of the item level information). Finally, five covariates were used for the ability parameter: gender, LCME[1] accreditation status, Native English speaker status, item response time, and ethnicity. Dummy variables were created to code ethnicity for 5 groups: Asian, Hispanic, Black, White and Other. White group was coded as the base group (i.e., indicated by the absence of coded ethnicity). Table 3.1 displays the percentages of examinees within each subgroup.

3.1.3 PTGI Data

The PTGI identifies five sub-domains of posttraumatic growth: Relating to Others, New Possibilities, Personal Strength, Spiritual Change, and Appreciation of Life. The survey comprises 21 five-point Likert-type items, all nested in five sub-domains of posttraumatic growth (i.e., testlets). The 718 participants participated in the study had recently been diagnosed with breast cancer and were administered the PTGI during an interview.

---

[1] The Liaison Committee on Medical Education (LCME) is the accrediting authority for medical education programs leading to the M.D. degree in U.S. and Canadian medical schools.

In addition to the 21 PTGI survey items, subjects were also asked background questions asked about race (White or not), ethnicity (Hispanic or not), age, income, employment status, marital status, whether or not they were taking Tamoxifen (a drug commonly used to assist in the prevention and recurrence of breast cancer in women near or beyond menopause), and how long it had been since they were diagnosed. Only θ-parameter covariates were used in the analyses because no variability in the item or testlet characteristics was observed. Table 3.2 displays the percentages of participants within each subgroup.

3.2 Parameter Estimation

All Bayesian calibrations that were carried out for this study were done using SCORIGHT (Wang, Bradlow, & Wainer, 2004). The default priors and hyperpriors specified in SCORIGHT were used in all three studies (i.e., simulation, USMLE, and PTGI).

3.2.1 Prior Distributions for Model Parameters

In a fully Bayesian framework, prior distributions are specified for every model parameter. The default priors in SCORIGHT for $\Lambda_1 = \{h_j, b_j, q_j, \theta_i, \gamma_{id(j)}\}$ were used in estimation. For the 2PL, priors for the item parameters, a_j and b_j are specified:

$$\begin{pmatrix} h_j \\ b_j \end{pmatrix} \sim \mathrm{N}_2\left(\begin{pmatrix} X_j^h \beta_h^{(2)} \\ X_j^b \beta_b^{(2)} \end{pmatrix}, \begin{pmatrix} (\sigma_h^{(2)})^2 & \rho_{hb}^{(2)} \sigma_h^{(2)} \sigma_b^{(2)} \\ \rho_{hb}^{(2)} \sigma_h^{(2)} \sigma_b^{(2)} & (\sigma_b^{(2)})^2 \end{pmatrix}\right) = N(\mu_{2PL}, \Sigma_{2PL}), \quad (3.1)$$

where $h_j = \log(a_j)$ and X_j^h and X_j^b are the covariates, and β_h and β_b are the slopes associated with *a* and *b*-parameters, respectively. Similarly, for the 3PL item parameter prior distributions a_j, and b_j, and c_j, are specified:

$$\begin{pmatrix} h_j \\ b_j \\ q_j \end{pmatrix} \sim \mathrm{N}_3 \left(\begin{pmatrix} X_j^h \beta_h^{(3)} \\ X_j^b \beta_b^{(3)} \\ X_j^q \beta_q^{(3)} \end{pmatrix}, \begin{pmatrix} (\sigma_h^{(3)})^2 & \rho_{hb}^{(3)} \sigma_h^{(3)} \sigma_b^{(3)} & \rho_{hq}^{(3)} \sigma_h^{(3)} \sigma_q^{(3)} \\ \rho_{hb}^{(3)} \sigma_h^{(3)} \sigma_b^{(3)} & (\sigma_b^{(3)})^2 & \rho_{bq}^{(3)} \sigma_b^{(3)} \sigma_q^{(3)} \\ \rho_{hq}^{(3)} \sigma_h^{(3)} \sigma_q^{(3)} & \rho_{bq}^{(3)} \sigma_b^{(3)} \sigma_q^{(3)} & (\sigma_q^{(3)})^2 \end{pmatrix} \right) = N(\mu_{3PL}, \Sigma_{3PL}), \quad (3.2)$$

where $q_j = \log(c_j / (1 - c_j))$ and X_j^q is the covariate, and β_q is the slope associated with *c*-parameter. Finally, the prior distributions for the polytomous item parameters are:

$$\begin{pmatrix} h_i \\ b_i \end{pmatrix} \sim \mathrm{N}_2 \left(\begin{pmatrix} X_i^h \beta_h^{(p)} \\ X_i^b \beta_b^{(p)} \end{pmatrix}, \begin{pmatrix} (\sigma_h^{(p)})^2 & \rho_{hb}^{(p)} \sigma_h^{(p)} \sigma_b^{(p)} \\ \rho_{hb}^{(p)} \sigma_h^{(p)} \sigma_b^{(p)} & (\sigma_b^{(p)})^2 \end{pmatrix} \right) = N(\mu_{poly}, \Sigma_{poly}). \quad (3.3)$$

The prior for the ability parameter was:

$$\theta_i \sim N(W_i \lambda, 1), \quad (3.4)$$

where λ is the covariate associated with the ability parameter. And the prior for the testlet parameter was:

$$\gamma_{id(j)} \sim N(0, \sigma_{d(j)}^2). \quad (3.5)$$

(Wang, Bradlow, & Wainer, 2005, p.4)

3.2.2 Hyperprior Distributions

In a fully Bayesian framework, hyperprior distributions are also specified for parameter prior distributions to reflect the uncertainty of their values. Wang et al. (2005) chose the hyperpriors to be conjugate and proper to the prior distributions. Hyperpriors for the following prior parameters were specified:

$$\Lambda_2 = \left\{ \lambda, \beta_h^{(3)}, \beta_b^{(3)}, \beta_q^{(3)}, \Sigma_{3PL}, \beta_h^{(2)}, \beta_b^{(2)}, \Sigma_{2PL}, \beta_h^{(p)}, \beta_b^{(p)}, \Sigma_{Poly}, \sigma_{d(j)}^2 \right\}.$$

The hyperprior for the ability parameter covariate coefficient was specified as:

$\lambda \sim N(0, \sigma_\lambda^2 \mathrm{I}_\mathrm{m})$, where $\sigma_\lambda^2 = 5$ and I_m is an *m*-dimensional identity matrix.

The hyperpriors for the 3PL item parameter covariate coefficients were specified as:

$$\beta_h^{(3)} \sim MVN(0, V_a),$$

$$\beta_b^{(3)} \sim MVN(0, V_b),$$

$$\beta_q^{(3)} \sim MVN(0, V_q),$$

where $|V_a|^{-1} = |V_b|^{-1} = |V_c|^{-1}$ are set to 0 to be noninformative (Wang et al, 2005). The same distributions follow for the 2PL and polytomous coefficients. For the covariances of the priors, the following slightly informative inverse-Wishart hyperpriors are set:

$$\Sigma_{3PL} \sim Inv - Wishart(3, M_3^{-1}),$$

$$\Sigma_{2PL} \sim Inv - Wishart(2, M_2^{-1}), \text{ and}$$

$$\Sigma_{Poly} \sim Inv - Wishart(2, M_2^{-1}),$$

where

$$M_3 = \begin{pmatrix} \frac{1}{100} & 0 & 0 \\ 0 & \frac{1}{100} & 0 \\ 0 & 0 & \frac{1}{100} \end{pmatrix}, \text{ and}$$

$$M_2 = \begin{pmatrix} \frac{1}{100} & 0 \\ 0 & \frac{1}{100} \end{pmatrix}.$$

(Wang, Bradlow, & Wainer, 2005, p.4)

3.2.3 Parameter Estimation for Ordinary Least-Squares Linear Regression

All regression analyses to estimate the covariate coefficients were carried out using Statistical Package for the Social Sciences (SPSS) version 14 (2005). The default options (i.e., were used to run linear multiple regression on the data sets. The SPSS Regression output includes coefficient estimates, their standard errors and their significance at the specified alpha level, as well as the confidence intervals. The output also displays information about the variation accounted for by the specified model.

3.3 Data Analysis

In this section, the analyses for the simulation data and the evaluation criterion for the simulation results are described. Analogous sections for the empirical data sets follow.

3.3.1 Simulation Study

SCORIGHT (Wang, Bradlow, & Wainer, 2004) was used for estimating the parameters. Three MCMC chains with 30,000 iterations (draws from posterior distribution) in each chain were run. The initial 10,000 draws were discarded after which every 20th draw was recorded to avoid autocorrelation. SCORIGHT manual (Wang, et al., 2005) recommends using Gelman and Rubin's (1993) criteria of $\sqrt{\hat{R}}_{.975} < 1.2$ for evaluating convergence (as described in Chapter 2). Using this as the criteria, convergence was observed across chains within all 100 runs.

SCORIGHT output includes posterior draws for all covariate coefficients. Using these draws, the Bayesian method was evaluated by examining the Coverage Probability (CP) criterion for each decile. For example, $CP_{90\%}$ is computed as follows:

$$CP_{90\%} = \frac{\sum_{i=0}^{n} 1(\text{true value} \in (\hat{F}^{(i)}_{0.95}, \hat{F}^{(i)}_{0.05}))}{n}, \qquad (3.6)$$

where $\hat{F}^{(i)}$ is the empirical cumulative distribution function of for posterior *i*, and *n* is the number of posterior distributions included in the analysis (Bradlow et al., 1999). Here, the expected coverage probability is .90; that is, the expected proportion of 90% credible intervals (i.e., the middle 90% of the empirical cumulative distribution function) that contain the true parameter values is .90. The observed coverage probability is the proportion of times the true value of the coefficient was within the observed 90% credible intervals.

To calculate the observed coverage probabilities for each decile, the posterior draws of the covariate coefficients for each replication were first sorted from low to high. The credible interval was then defined as the interval containing the middle *n*% of draws, where *n* corresponds to the interval size (e.g., 90%). For each interval, the proportion of replications for which the true covariate fell in the credible interval was taken as the observed coverage probability. Across the 50 replications and for each covariate coefficient, this was computed as follows:

$$CP_{\beta,90\%} = \frac{\sum_{i=1}^{50} \beta \in (\hat{F}^{(i)}_{\beta,0.05}, \hat{F}^{(i)}_{\beta,0.95})}{50}. \quad (3.7)$$

When $CP_{\beta,90\%}$ differs from 90%, it is due to sampling error, estimation error, or both.

To evaluate the post-hoc regression approach, confidence intervals for each centered decile were computed for each replication's covariate coefficients using the estimates of the covariates and their respective standard errors. Next, the proportion of times the true covariate fell within the given confidence interval recorded. This observed proportion was then compared with the expectation. For example, for a 90% confidence interval, the proportion of times the true covariate fell within the 90% confidence interval for each estimated covariate was then compared to an expectation of .90.

In the frequentist framework, a 90% confidence interval means that 90% of the confidence intervals would contain the true parameter were the analysis repeated a large number of times for different samples (drawn from the same population). Given a single sample, the outcome that the true parameter lays within the calculated confidence interval

is either 0 or 1. However, when this is repeated 50 times, as it was in this study, the proportion of times the parameter lies in the confidence interval can be computed and the concept of confidence interval becomes comparable to the concept of Bayesian coverage probability. Figure 1 illustrates how this concept could be similar to Bayesian coverage probability with multiple replications for parameter μ. In this way, the relative success of the Bayesian approach and post-hoc regression approach could be contrasted.

Both approaches were also examined in terms of bias and Root Mean Square Error (RMSE). Bias was determined by computing the signed difference between the true parameter and the estimate and then averaging these signed differences across 50 replications. Equation 3.8 illustrates the bias calculation for the regression coefficient.

$$Bias(\hat{\beta}_j) = \frac{\sum_{r=1}^{N}(\hat{\beta}_{jr} - \beta_j)}{N} \tag{3.8}$$

where $\hat{\beta}_{jr}$ is the estimated coefficient for item covariate j in replication r.

RMSE, as shown in equation 3.9, was also computed for each coefficient estimate.

$$RMSE(\hat{\beta}_j) = \sqrt{\frac{\sum_{r=1}^{N}(\hat{\beta}_{jr} - \beta_j)^2}{N}} \tag{3.9}$$

where $\hat{\beta}_{jr}$ is the estimated coefficient for item covariate j in replication r.

3.3.2 USMLE Step 3 and PTGI Data

USMLE Step 3 data were analyzed using the Bayesian TRT model to simultaneously estimate each USMLE Step 3 examinee's ability, the item parameters, the

testlet structure of the test, and the covariate coefficients. Again, SCORIGHT (Wang, Bradlow, & Wainer, 2004) was used for estimation. All items were multiple-choice and therefore, the 2PL TRT model was used to calibrate the data. Three MCMC chains with 30,000 iterations per each chain were run. The initial 10,000 draws were discarded after which every 20^{th} draw was recorded.

The PTGI was designed to assess changes in life following a traumatic event. Within the TRT framework, this was accomplished by estimating the location of the respondents on the latent dimension of interest where those with a higher value on the latent dimension are more likely to give higher ordinal response scores. The model that SCORIGHT uses for analyzing polytomous data is Samejima's GRM (1969). Three MCMC chains were run with 30,000 draws in each. The initial 10,000 draws were again discarded after which every 20^{th} draw was recorded.

Since all items are nested within testlets in PTGI, in addition to modeling parameters, the local dependence problem and its impact on the inferences for the PTGI data was also investigated.

Since the purpose of the empirical data analyses was to compare results with the conventional post-hoc approach, conventional regression analyses were conducted for both data sets, using the model parameter point estimates as the dependent variables. Note that these point estimates were obtained from a SCORIGHT run that did not include covariates. This was necessary because when covariates are present in the model, their coefficients are simultaneously estimated along with the model parameters.

Three pieces of information were used to interpret the Bayesian results. SCORIGHT provides draws from the posterior distribution for each covariate coefficient

as well as the mean and standard deviation of these draws. In addition, using the posterior draws for each covariate, the proportion of draws larger or smaller than zero for each coefficient was calculated. To visually examine these proportions, the kernel densities of the posterior draws were also plotted. To interpret the post-hoc regression results, significance of the estimates were considered and confidence interval (CI) for each coefficient was calculated. However, recall from section 3.3.1 above that strictly speaking, confidence intervals and credible intervals, while analogous, are not identical concepts particularly for a single occasion (as opposed to many replications). This difference highlights an important point.

The interest here is in whether or not these two approaches to modeling covariates lead to different inferences about the relationships they describe. Of particular interest are differences that arise due to the failure of the regression approach to account the uncertainty in the response model parameters. It is hoped that the simulation study described above will provide some indication of the nature of such differences and the results from the empirical study will be scrutinized from this perspective. However, in addition, the form of the covariate estimates associated with the two approaches (posterior draws for the Bayesian model and point estimates with standard errors for the regression approach) may also lead to different inferences. These differences, should any be observed, will also be discussed in Chapter 5.

Table 3.1. True Regression Coefficients

	Beta 1	Beta 2
Condition 1	0.482853	0.52021
Condition 2	0.339836	0.08825

Figure 3.1. Representation of a 95%-Confidence Interval across 50 Samples

CHAPTER 4

RESULTS

This chapter is divided into three main sections: simulation study results, USMLE Step 3 data set results, and PTGI data set results. After reviewing the detailed results, the implications will be discussed in chapter 5.

4.1 Simulation Study Results

The main purpose of the simulation study was to examine the accuracy and effectiveness of the Bayesian TRT approach in estimating the relationship of covariates to their respective model parameters. This approach was also compared to the conventional post-hoc regression approach. As described in chapter 3, the scope of the simulation study was limited to proficiency parameter covariates. In this section, analyses of these data are described in detail for both simulation conditions.

As an aside, prior to running the post-hoc regression analyses, the thetas were transformed onto the true theta metric via mean-sigma equating. This was important so that the resulting regression coefficients would be on the same scale as the true coefficients. Similarly, the Bayesian coefficient estimates were also rescaled on to the true metric by multiplying each coefficient by the true thetas' standard deviation and dividing it by the estimated thetas' standard deviation.

4.1.1 Post-hoc Regression Approach

As explained in chapter 3, for the post-hoc analyses, SCORIGHT calibrations were run without covariates in the model to obtain theta estimates. All replications converged according to Gelman and Rubin's (1993) criteria of $\sqrt{\hat{R}}_{.975} < 1.2$ (as described in chapter 2). First, the correlation between the generating thetas and the estimated thetas for each replication was examined. The average correlation across 50 replications for Condition 1 was .98 (SD=.0008). Next, the correlations between the covariates and the estimated thetas were compared to the correlations between the covariates and the true theta for Condition 1. The observed correlations were lower than the true correlations on average with the mean difference for both covariates being approximately .0.02. The correlation between the covariates and true theta was always underestimated by the correlation between the estimated thetas and the covariates, albeit only by small amounts.

The average correlation between the estimated and true thetas across the 50 replications for Condition 2 was .97 (SD=.0009). As with Condition 1, the Condition 2 correlations between covariates and the estimated thetas were always slightly lower than the correlations between true thetas, with a mean difference of 0.01 for both covariates.

To evaluate the relative success of the post-hoc regression approach, confidence intervals for each centered decile were computed for each replication's covariate coefficient using the estimates of the covariate coefficients and their respective standard errors. Next, the proportion of times the true covariate fell within the given confidence interval recorded. This observed proportion was then compared with expectation. For example, for a 90% confidence interval, the proportion of times the true covariate fell

within the 90% confidence interval for each estimated covariate was then compared to an expectation of .90.

Figures 4.1 to 4.9 show the covariate confidence intervals for the centered deciles for Condition 1. It is apparent from these figures that the post-hoc approach consistently underestimated the covariate coefficients for Condition 1. Moreover, for the first 7 confidence intervals, the true parameter fell within the interval fewer times than expected. This result is displayed in Figures 4.37 and 4.38.

Similar to Condition 1, Figures 4.10 to 4.18 show the covariate confidence intervals for the centered deciles for Condition 2. Again, the post-hoc approach resulted in a general underestimation trend of the covariate coefficients. With respect to the confidence intervals, however, Condition 2, in contrast to Condition 1, tended to yield intervals containing the true parameter more often than expectation. This finding is shown in Figures 4.39 and 4.40.

4.1.2 Bayesian Approach

Using the covariate coefficients' posterior draws obtained from SCORIGHT, the Bayesian method was evaluated by examining the credible intervals and coverage probabilities for each decile. This is analogous to the analysis of the confidence intervals for the post-hoc approach. As with the non-covariate calibrations described above, convergence was observed for all replications according to Gelman and Rubin's (1993) criterion of $\sqrt{\hat{R}}_{.975} < 1.2$.

As with the post-hoc analyses, the correlations between the generating thetas and the estimated thetas from the runs with covariates were examined for Condition 1. The

average correlation across 50 replications was 0.97 (SD= .001). Next, the correlations between the covariates and the estimated thetas were compared to the correlations between the true thetas and the covariates for Condition 1. In contrast to the post-hoc results, the observed correlations were always higher than the true correlations on average with a mean absolute difference of .02 for both covariates.

The same analyses were done for Condition 2. Here, the correlation between the generating thetas and the estimated thetas for each Condition 2 replication was examined and it was found that the average correlation across replications was 0.97 (SD= .001). Next, the correlations between the covariates and the estimated thetas were compared to the correlations between the true thetas and covariates for Condition 2. As with Condition 1, the observed correlations were always higher than the true correlations with a mean absolute difference for Covariate 1 of .008 and for Covariate 2 of .002. In other words, the correlations between the covariates and estimated theta were consistently, albeit slightly, overestimated across replications.

Across 50 replications for both conditions, the expected coverage probability was compared to the observed coverage probability for each centered decile (credible interval). For each interval, the proportion of replications for which the true covariate fell in the credible interval was taken as the observed coverage probability and as with the confidence intervals, the expectation was the decile percentage (e.g., for a 90% credible interval, the expected coverage probability is .90).

As expected, the Bayesian approach resulted in relatively wider credible intervals than the confidence intervals produced by the post hoc-approach. For example, when averaged across the two covariates, the estimated interval for 90% was 0.005 wider for

the Bayesian model than the post-hoc-model for Condition 1. Figures 4.19 to 4.27 show the credible intervals for each replication and decile, revealing that the Bayesian approach slightly overestimates the betas. Figures 4.37 and 4.38 show that the Bayesian approach was more accurate than the post-hoc approach in terms of recovering the expected coverage probability for each credible interval up to the 80% interval for Covariate 1, and 70% interval for Covariate 2. For larger intervals, which are typically of greater interest, the trend changes and post-hoc approach is slightly superior.

For Condition 2, the expected coverage probability was also compared to the observed coverage probability for each decile. Again, the Bayesian approach resulted in slightly wider intervals than the post hoc-approach. For example, when averaged across two covariates, the estimated interval for 90% was 0.002 wider for the Bayesian model than the post-hoc-model. As with Condition 1, Figures 4.28 to 4.36 suggest that the Bayesian approach tends to slightly overestimate the betas. Figures 4.39 and 4.40 show that the Bayesian approach and the post-hoc approach performed almost identically with respect to recovering the expected coverage probability for each credible interval except the 10% and 20% intervals for Covariate 1 where the post-hoc approach was closer to expectation.

4.1.3 RMSE and Bias

The RMSE and bias of the covariate coefficients was computed and also averaged (*absolute* bias was averaged) across replications for the both post-hoc regression and the Bayesian approaches. These values are displayed in Table 4.1 for Condition 1 and 2. Results at the covariate level were mixed. For Condition 1, the Bayesian approach was

superior on Covariate 1 ($bias_{Bayesian}$ = 0.010 vs. $bias_{post\text{-}hoc}$ = -.015; $RMSE_{Bayesian}$ =0.011 vs. $RMSE_{post\text{-}hoc}$ = 0.016), and inferior on Covariate 2 ($bias_{Bayesian}$ = -0.014 vs. $bias_{post\text{-}hoc}$ = 0.012; $RMSE_{Bayesian}$ = 0.015 vs. $RMSE_{post\text{-}hoc}$ =0.013).

For Condition 2, the Bayesian approach was superior with respect to Covariate 2 bias ($bias_{Bayesian}$ = -0.001 vs. $bias_{post\text{-}hoc}$ = 0.004), and RMSE ($RMSE_{Bayesian}$ = 0.005 vs. $RMSE_{post\text{-}hoc}$ =0.006), but performed the same with respect to RMSE and absolute bias for Covariate 1 (0.010 and 0.009, respectively). When averaged across covariates and conditions, the Bayesian approach had slightly less absolute bias ($bias_{Bayesian}$ = 0.009 vs. $bias_{post\text{-}hoc}$ = 0.010) and the same RMSE (0.01). Nevertheless, these differences are extremely small in almost all cases, indicating that the Bayesian and post-hoc approaches are performing very similarly in terms of capturing the covariate slopes.

4.2 USMLE Step 3 Results

The convergence statistics confirmed proper convergence after 10,000 initial iterations according to Gelman and Rubin's (1993) criterion of $\sqrt{\hat{R}}_{.975} < 1.2$. The examinee covariate distribution is displayed in Table 4.2. The Bayesian posterior estimates of the item and testlet parameter covariate coefficients and their respective standard errors are given in Table 4.3. Figures 4.41-4.49 display the kernel densities of the coefficients of the item and testlet parameter covariates.

As the figures and the coefficients reveal, the covariates associated with the discrimination, difficulty, and testlet effect parameters are all near zero, indicating that *a*-, *b*-, and γ -parameters may not be significantly related to their covariates: *vignette word count*, *stem word count*, or *options word count*. This assertion may be tested by

constructing credible intervals. Here, the value of interest is zero and the interval of interest is either an uppermost or lowermost region, analogous to a one-tailed test. Thus, a very simple way to evaluate an estimate is to merely count the number of times the posterior draw is greater (or less) than zero. In so doing, the empirical cumulative distribution function described by the posterior draws is taken to be proportional to the probability density function for beta and thus, the proportion of draws greater (or less) than zero may be interpreted as the probability that beta is greater (or less) than zero. The *counting* approach is analogous to a one-tailed hypothesis test with the advantage of yielding a posterior probability, which may be more intuitively understood than frequentist results. In any case, to compare the Bayesian results to the post-hoc results, the approach taken here was to examinee whether the upper (or lower, if the covariate coefficient estimate is negative) 95% of the posterior distribution includes zero or not. Although the underlying perspective of the frequentist approach differs, this criterion is somewhat comparable to the post-hoc criterion of α =.05. In fact, in the case of repeated sampling and uninformative priors for the Bayesian approach, if both approaches were equally effective and error was random, we would expect that the two approaches would yield the same conclusions.

For example, consider the covariate *stem word count*. Here, the mean of the posteriors is in both cases positive and therefore, the probabilities that the coefficients are greater than zero are computed. The proportions of draws from the regression weight's posteriors that were greater than zero for the *a*- and *b*-parameters were .44 and .49, respectively. Thus, the posterior probabilities that items with relatively more words in their stem will be more discriminating and difficult are .44 and .49, respectively, both of

which are below the criterion of .95. The post-hoc results also support the same conclusion that these slopes are not likely to be different than zero, with a *p*-value of .45 for the *a*-parameter and .95 for the *b*-parameter.

The coefficient estimates for the covariates of the θ-parameter are given in Table 4.4 and Figures 4.50-4.57 display the kernel density plots for the person covariates. The coefficient values and the figures suggest that most person covariates have a significant relationship with the ability parameter. To investigate further, the probability of drawing a value greater (or less) than zero for each coefficient was calculated using the empirical cumulative distribution function described by the posterior draws. The results indicated a positive relationship between LCME status ($\bar{b} = 0.87$, and $P(b > 0) = 1.00$) and proficiency; and being a native English speaker ($\bar{b} = 0.47$, and $P(b > 0) = 1.00$) and proficiency. Response time also seems to play a role ($\bar{b} = -0.05$, and $P(b < 0) = 1.00$): the shorter the response time, the higher the estimated ability was. For each ethnicity/race subgroup, the probabilities of each group's coefficients being greater than (or less than) zero compared to the base group are also reported in Table 4.4. *White* group membership was positively related to proficiency ($\bar{b} = 0.39$, and $P(b > 0) = 0.98$), while *Black* group membership ($\bar{b} = -0.63$, and $P(b < 0) = 0.99$) was negatively related to proficiency. *Asian* ($\bar{b} = -0.15$, and $P(b < 0) = 0.78$) and *Hispanic* (($\bar{b} = -0.26$, and $P(b < 0) = 0.85$) group memberships were not significantly related to proficiency, given the .95 criterion. Each race/ethnicity group was also compared to one another in Figures 4.58 to 4.63. These comparisons give further information on specific group differences. The corresponding probabilities for the differences are given in Table 4.5. The *p*-values for comparing the *White* group the other race/ethnicity groups were all 1, indicating

significant differences; that is, *White* group membership was more positively related to proficiency than the other three groups. The *p*-value for comparing the *Asian* to the *Black* group was .98, again indicating a significant difference favoring the *Asian* group. The comparison between *Hispanic* and *Black* groups (p=.91) and *Asian* and *Hispanic* groups (p=.71) did not show significant differences.

Conventional regression analyses using item and person parameters as dependent variables were also conducted. The results indicated that none of the covariates for any of the item parameters were significant, confirming the Bayesian interpretation (Tables 4.6 and 4.7). The results for the proficiency parameter were again in line with the Bayesian interpretation (Table 4.8); LCME status, native English Speaker, Response Time, and *White* and *Black* group membership were all significantly related to the ability parameter.

4.3 PTGI Survey Results

The convergence statistics indicated proper convergence after the burn-in period according to Gelman and Rubin's (1993) criterion of $\sqrt{\hat{R}}_{.975} < 1.2$. The survey participants' covariate distributions are displayed in table 4.9. The Bayesian posterior estimates of the person parameter covariates, their respective standard errors, and the probability of drawing a value larger than zero for each covariate coefficient are given in Table 4.10. Figures 4.64-4.71 display the kernel densities of the coefficients of the covariates.

The figures and the covariate coefficient estimates indicate that a number of covariates have a significant relationship with the θ-parameter, which in this case is the latent posttraumatic growth and will be referred to as *changeability* from this point on.

The results show that use of *Tamoxifen* ($\bar{b} = 0.32$, and $P(b > 0) = 1.00$) and *Work Status* (being employed or not) ($\bar{b} = 0.18$, and $P(b > 0) = 0.97$) are both positively related to changeability while *White* group membership is negatively related to changeability ($\bar{b} = -0.21$, and $P(b < 0) = 0.99$). *Age* ($\bar{b} = -0.03$, $P(b < 0) = 1.00$) and *Months Since Diagnosis* ($\bar{b} = -0.03$, $P(b < 0) = 0.97$) are also negatively related to changeability. The covariates *Hispanic* group membership, *Income*, and *Married* were excluded because they failed to satisfy the .95 criterion: $P(b > 0)$ was 0.92, 0.71 and 0.36, respectively.

The results of conventional post-hoc regression analyses using theta as the dependent variable also confirmed the Bayesian interpretation that use of *Tamoxifen and Age* are both significant (Table 4.11). However, post-hoc analyses failed to show significance of *Months Since Diagnosis*, *White* group membership and *Work Status* that the Bayesian analyses did. Moreover, the post-hoc analyses indicated a non-significant effect of *Hispanic* group membership (p=.24) but if the posterior distribution for the slopes (Figure 4.65) is examined, it is observed that 92% of the distribution lies below zero ($\bar{b} = 0.19$). Although .92 is also below the Bayesian criterion, unlike the post-hoc result, the Bayesian result is high enough to perhaps justify further inquiry such as collecting more data to increase power.

4.3.1 Effect of Local Dependence

The local dependence problem is not the main focus of this research; however, since all items are nested within testlets in PTGI, it seems investigating the effect of unmodeled dependence in the context of PTGI is warranted. The survey instrument was divided into five testlets, as previously described. This structure was empirically

confirmed by a principal components analysis performed by Tedeschi and Calhoun (1996) and used as supporting evidence for the format of the survey. The model that was fit allowed dependence within testlets by including testlet effects. If these same data were fit with the analogous IRT (not TRT) model that assumes local independence, for point estimates of the parameters, the results would not be very different. However, unmodeled local dependence may have an impact on the results through measurement precision (as explained in chapter 2). If independence is falsely assumed, one does not have as much information as it is believed. This was what was investigated in this section.

In Figure 4.72 the posterior distributions of the variance of testlet parameter γ is shown. If the σ_γ^2 is zero, there is no local dependence. The extent to which it is greater than zero is a measure of the local dependence. To norm the size of σ_γ^2, the ratio between σ_γ^2 and the variance of θ is normalized to 1 (Wainer, et al., 2007). Therefore, γ is on the same metric as θ. Table 4.12 contains the estimated variance of the testlet parameter for each testlet. As it can be seen in Figure 4.72 and Table 4.12, the testlet parameters' posteriors are greater than zero for all five testlets, but especially for Spiritual Change ($\hat{\sigma}_\gamma^2$=0.46).

There appears to be some local dependence within testlets. But how much does this dependence affect information if one fails to model it? To answer this question, the posterior distributions of the proficiency parameter estimate for a typical respondent are examined for two models that are identical except that one assumes local independence (IRT) and one does not (TRT). As is evident in Figure 4.73, both distributions were centered on the same point in the continuum, but the TRT model is more platykurtic. The

IRT based distribution indicates greater precision than there actually was. In fact, when averaged over all 718 respondents, the variance of the posterior was underestimated by about 50%. This means that there is, in fact, less information to estimate theta than IRT model would estimate.

Table 4.1. RMSE and Absolute Bias for the Two Approaches Averaged across Conditions and Covariates

Post-hoc		Bayesian	
RMSE	BIAS	RMSE	BIAS
0.011	0.010	0.010	0.009

Table 4.2. Examinee Covariate Distribution

Sex	
Male	56%
Female	44%
Race	
White	50%
Asian	30%
Hispanic	7 %
Other	7%
Black	6%
LCME status	
Yes	56%
No	44%
Native English Speaker	
Yes	58%
No	42%

Table 4.3. Estimated Coefficients of the Item Discrimination, Difficulty, and Testlet Parameter Covariates

	Covariate	Posterior Mean	Posterior std.	P(b>0)
a-parameter	Intercept	-1.70	0.24	-
	Vignette word count	0.00	0.00	0.88
	Stem word count	0.00	0.00	0.44
	Options word count	0.01	0.00	0.94
b-parameter	Intercept	-5.61	1.43	-
	Vignette word count	0.00	0.01	0.68
	Stem word count	0.00	0.01	0.49
	Options word count	0.02	0.02	0.83
γ-parameter	Intercept	1.37	1.48	-
	Vignette word count	0.00	0.00	0.65
	Stem word count	-0.01	0.01	0.75
	Options word count	0.00	0.00	0.67

Table 4.4. Estimated Coefficients of Theta Parameter Covariates

Covariate	Posterior Mean	Posterior std.	P(b>0)
Gender	0.13	0.10	0.90
LCME status	0.87	0.13	1.00
Native English speaker	0.47	0.13	1.00
Item response time	-0.05	0.01	0.00
White	0.39	0.19	0.98
Asian	-0.15	0.19	0.22
Black	-0.63	0.26	0.01
Hispanic	-0.26	0.24	0.15

Table 4.5. P-values for the Differences of Proficiency between Four Racial/Ethnic Groups

Race/Ethnicity	*p*-value
White-Hispanic	1.00
White-Asian	1.00
White-Black	1.00
Asian-Hispanic	0.71
Asian-Black	0.98
Hispanic-Black	0.91

Table 4.6. Results of Post-hoc Regression Analyses for a-parameter Covariates

Covariate	Coefficient	S.E.	Probability
Vignette word count	0.00	0.00	0.38
Stem word count	0.00	0.00	0.45
Options word count	0.00	0.00	0.27

Table 4.7. Results of Post-hoc Regression Analyses for b-parameter Covariates

Covariate	Coefficient	S.E.	Probability
Vignette word count	0.00	0.00	0.55
Stem word count	0.00	0.01	0.95
Options word count	0.01	0.01	0.45

Table 4.8. Results of Post-hoc Regression Analyses for θ-parameter Covariates

Covariate	Coefficient	S.E.	Probability
Gender	0.00	0.59	0.19
LCME status	0.54	0.08	0.00
Native English speaker	0.25	0.08	0.00
Asian	0.00	0.12	0.40
Hispanic	-0.20	0.15	0.18
Black	-0.43	0.16	0.01
White	0.24	0.12	0.04
Response time	0.00	0.04	0.00

Table 4.9. Participants' Person Covariate Distribution

Marital Status	
Married	57%
Race/Ethnicity	
White	70%
Hispanic	11%
Work Status	
Working	60%
Household Income	
<10K	7%
10K-20K	12%
20K-30K	13%
30K-50K	23%
>50K	29%
Currently taking Tamoxifen	
Yes	44%

Table 4.10. Estimated Coefficients of θ-Parameter Covariates

Covariate	Posterior Mean	Posterior std.	P(b>0)
Age	-0.03	0.00	0.00
Months since Diagnosis	-0.03	0.02	0.03
White	-0.21	0.09	0.01
Hispanic	0.19	0.13	0.92
Married	-0.03	0.09	0.36
Work Status (Working)	0.18	0.10	0.97
Income	0.02	0.04	0.71
Tamoxifen	0.32	0.08	1.00

Table 4.11. Results of Post-hoc Regression Analyses for θ-parameter Covariates

Covariate	Coefficient	S.E.	Probability
White	-0.16	0.10	0.11
Hispanic	0.13	0.13	0.32
Married	-0.06	0.10	0.51
Work Status (Working)	0.14	0.10	0.17
Income	0.03	0.04	0.43
Tamoxifen	0.33	0.08	0.00
Age	-0.03	0.01	0.00
Months since diagnosis	-0.03	0.02	0.06

Table 4.12. Estimated Variance of Gamma for Each Testlet

Testlet	σ_γ^2	S.E.
Personal Strength	0.14	0.02
Relating to Others	0.15	0.02
New Possibilities	0.18	0.03
Appreciation of Life	0.19	0.03
Spiritual Change	0.46	0.06

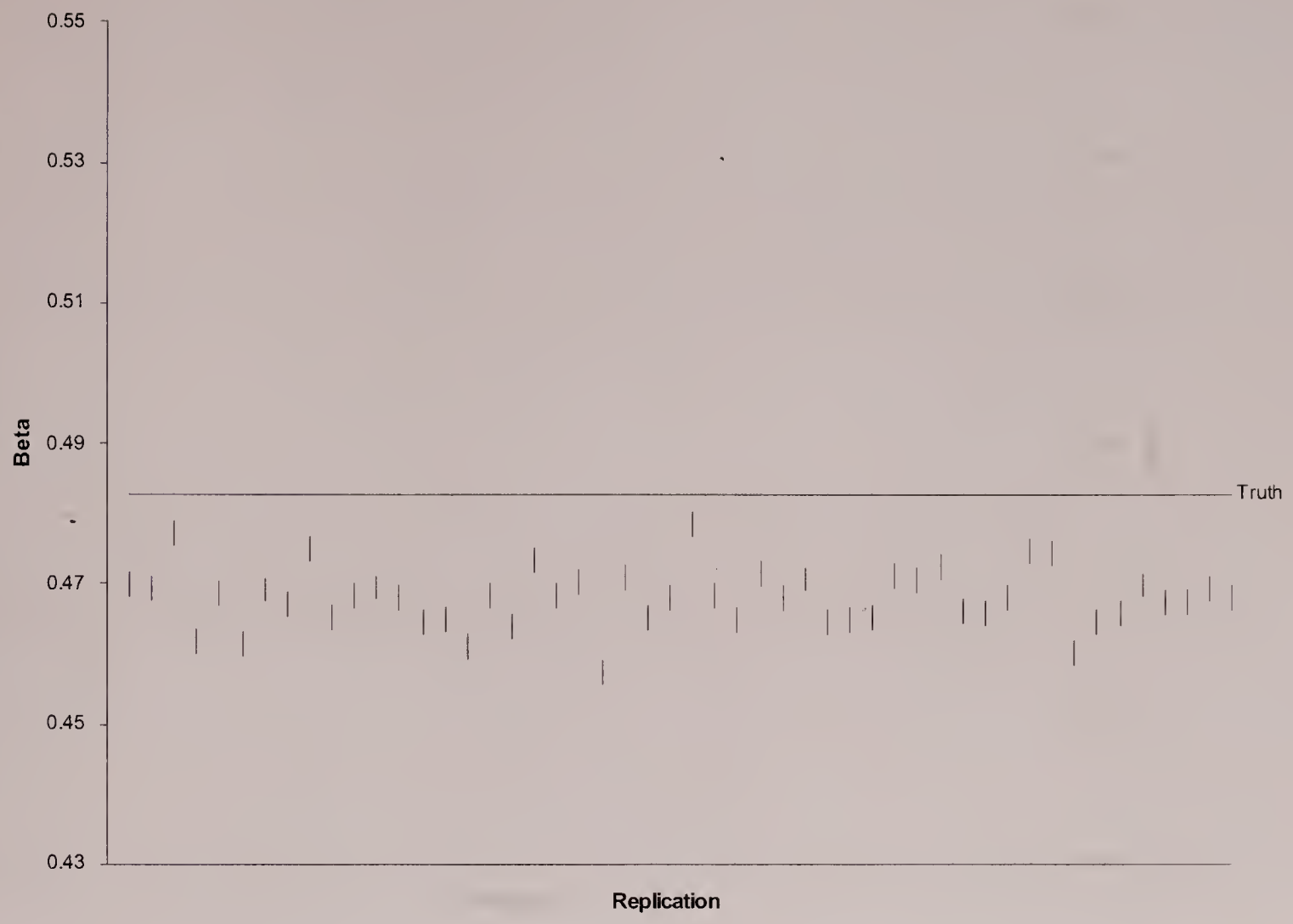

Figure 4.1. 10% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

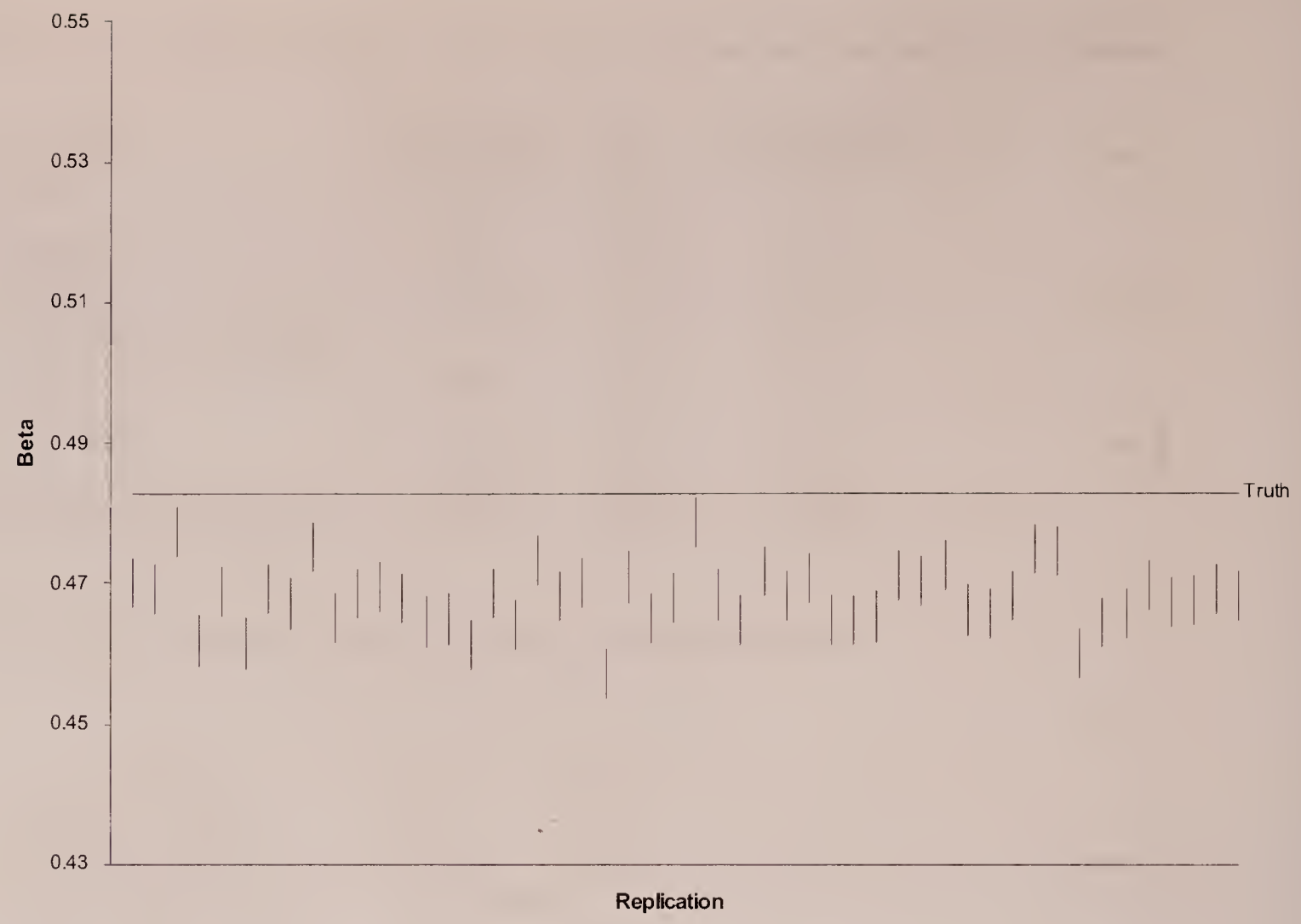

Figure 4.2. 20% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

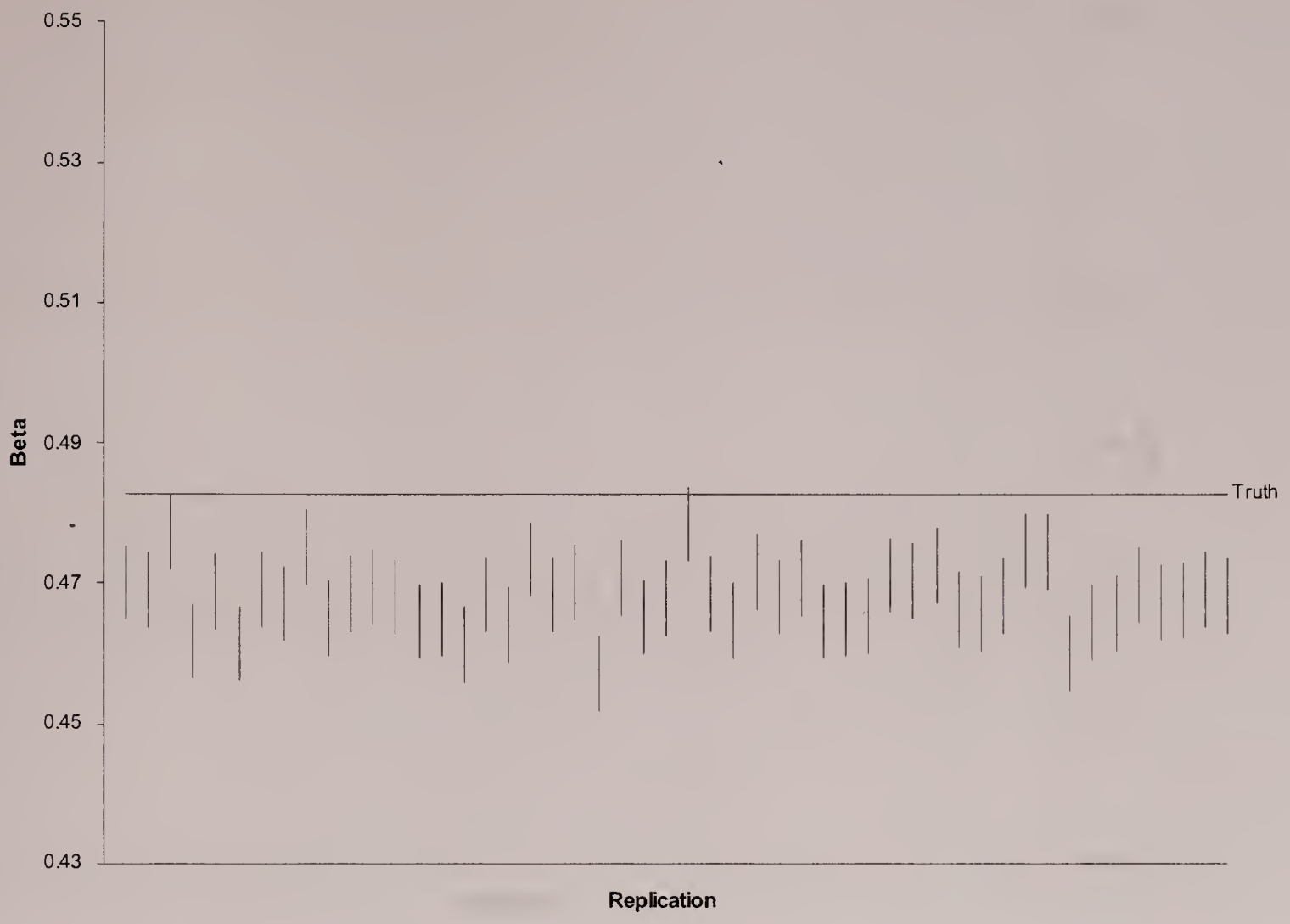

Figure 4.3. 30% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

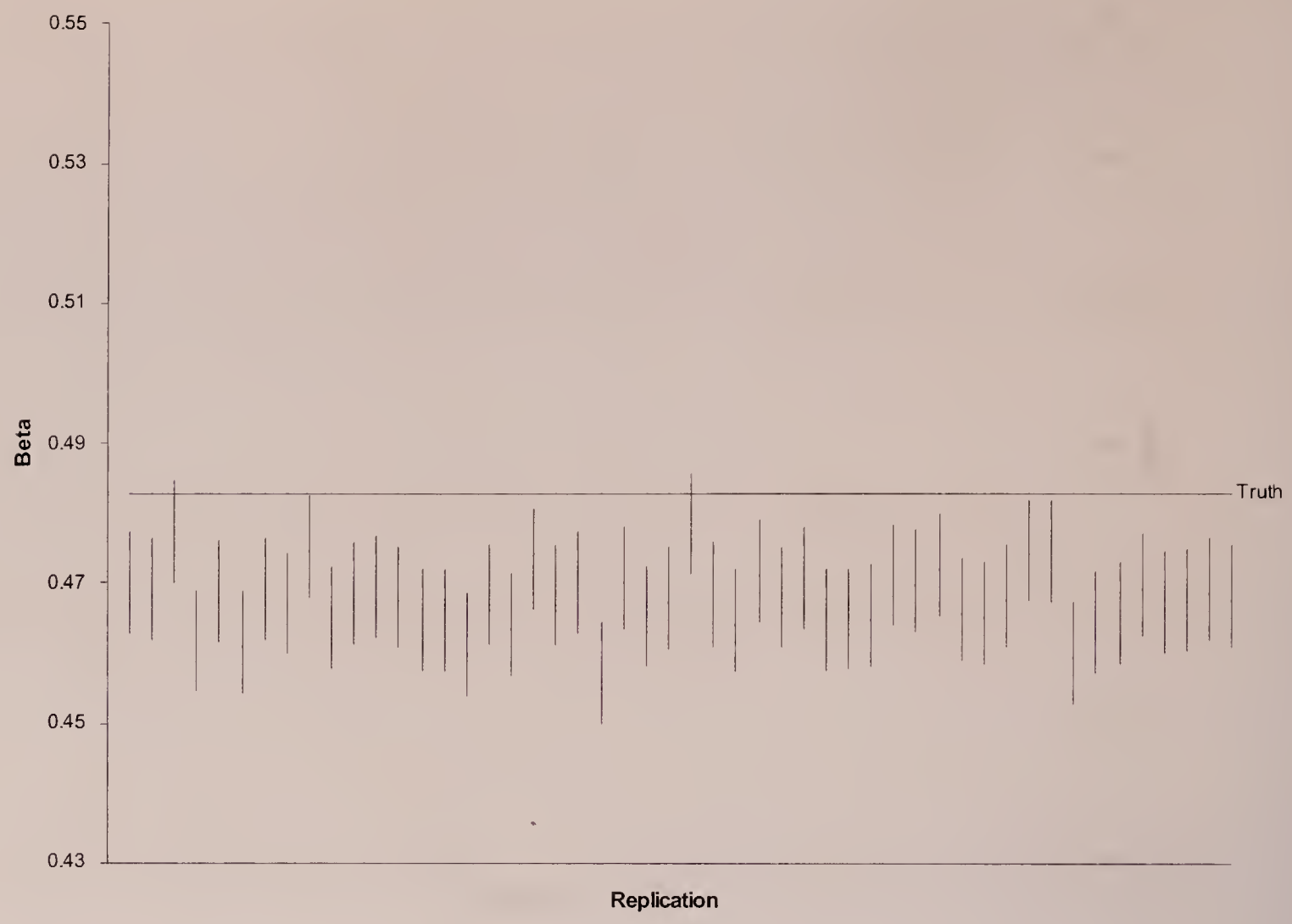

Figure 4.4. 40% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

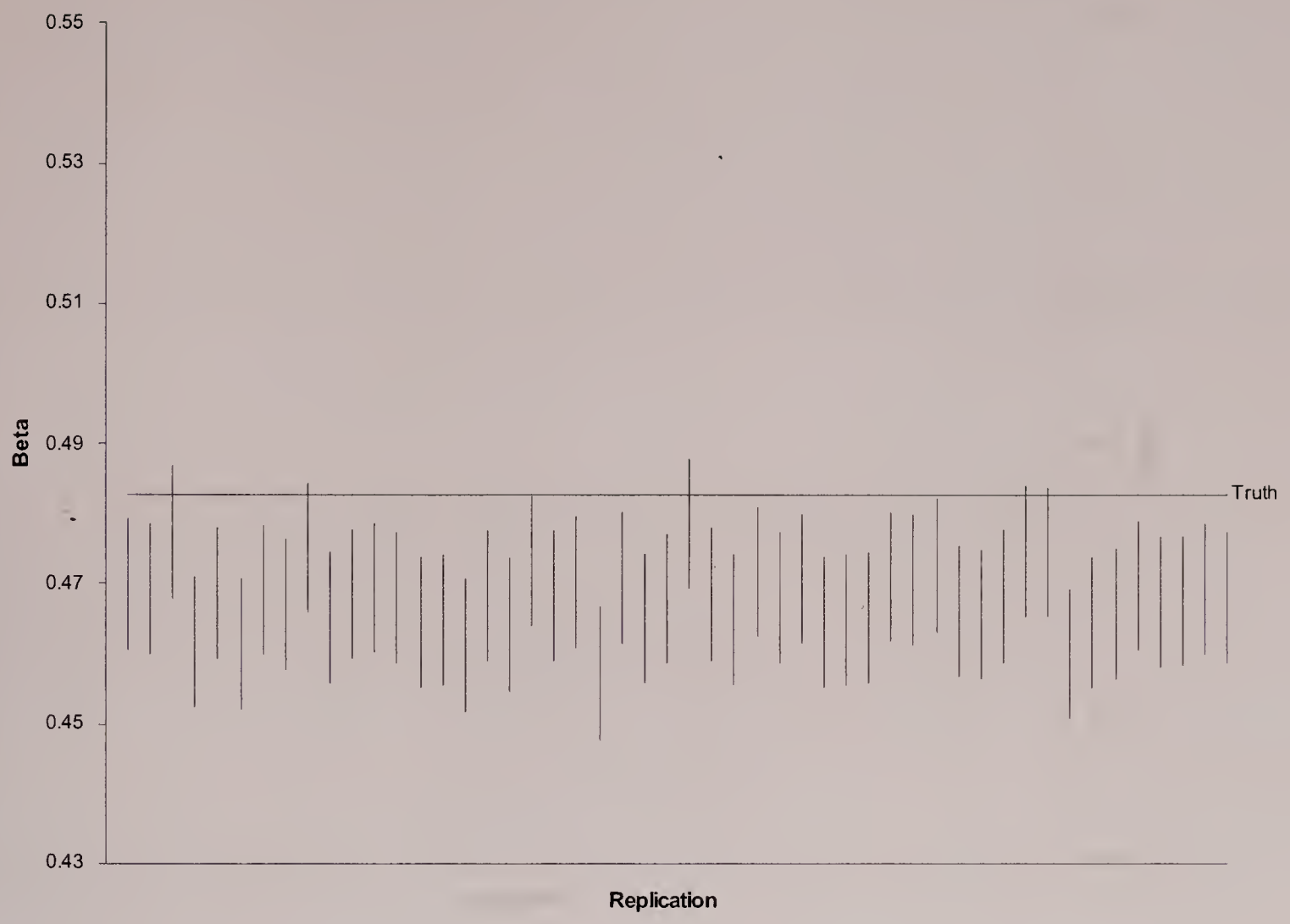

Figure 4.5. 50% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

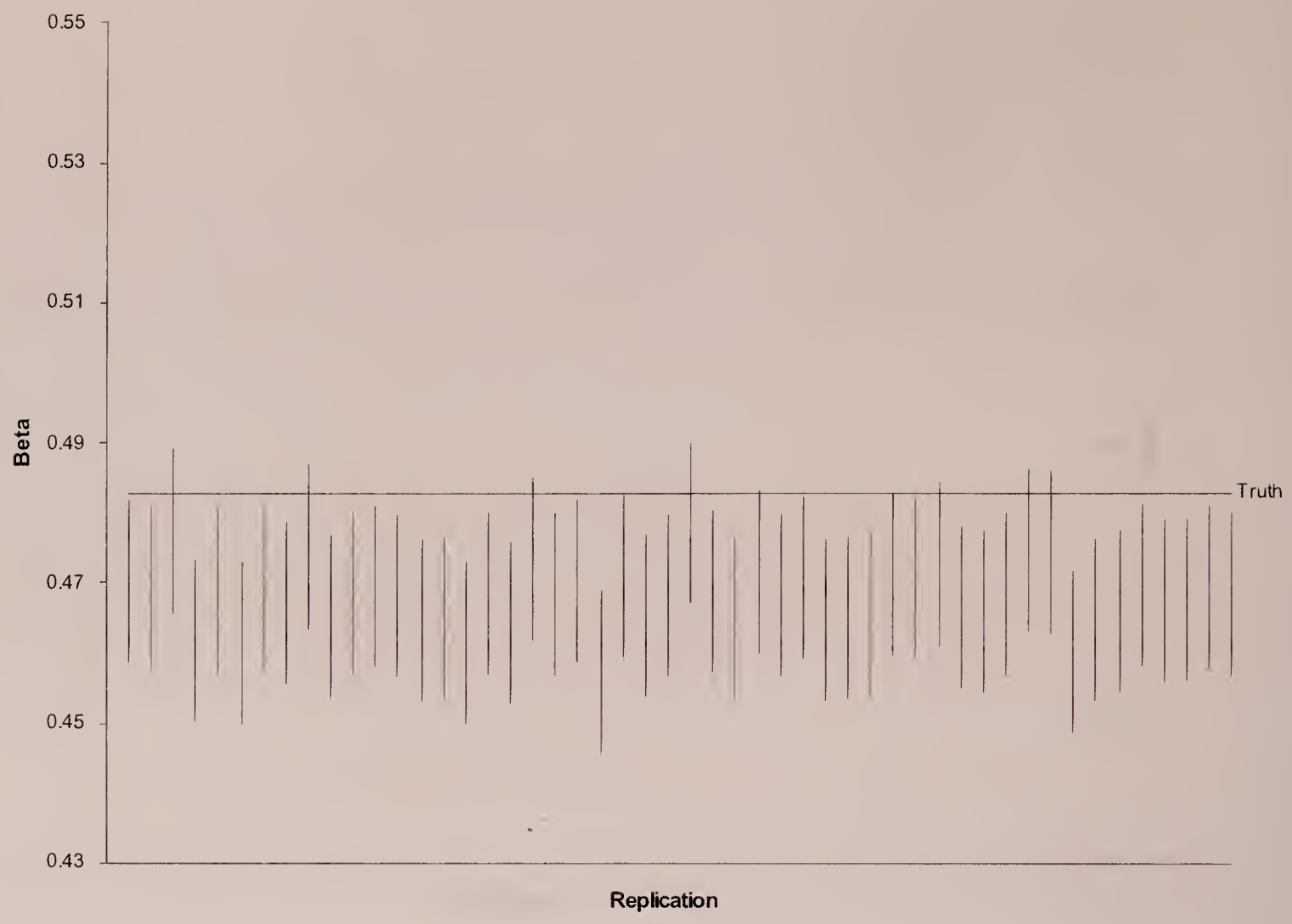

Figure 4.6. 60% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

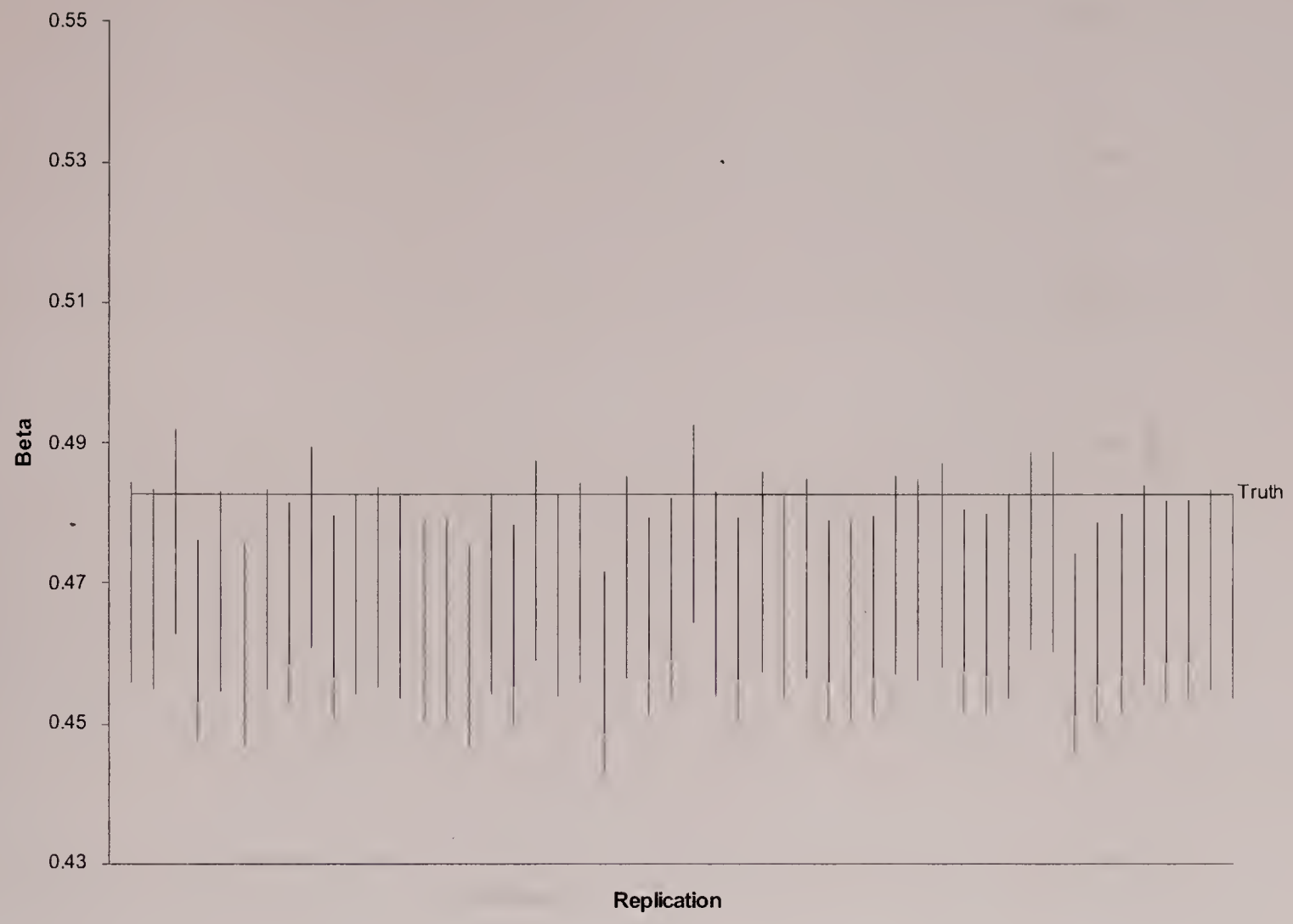

Figure 4.7. 70% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

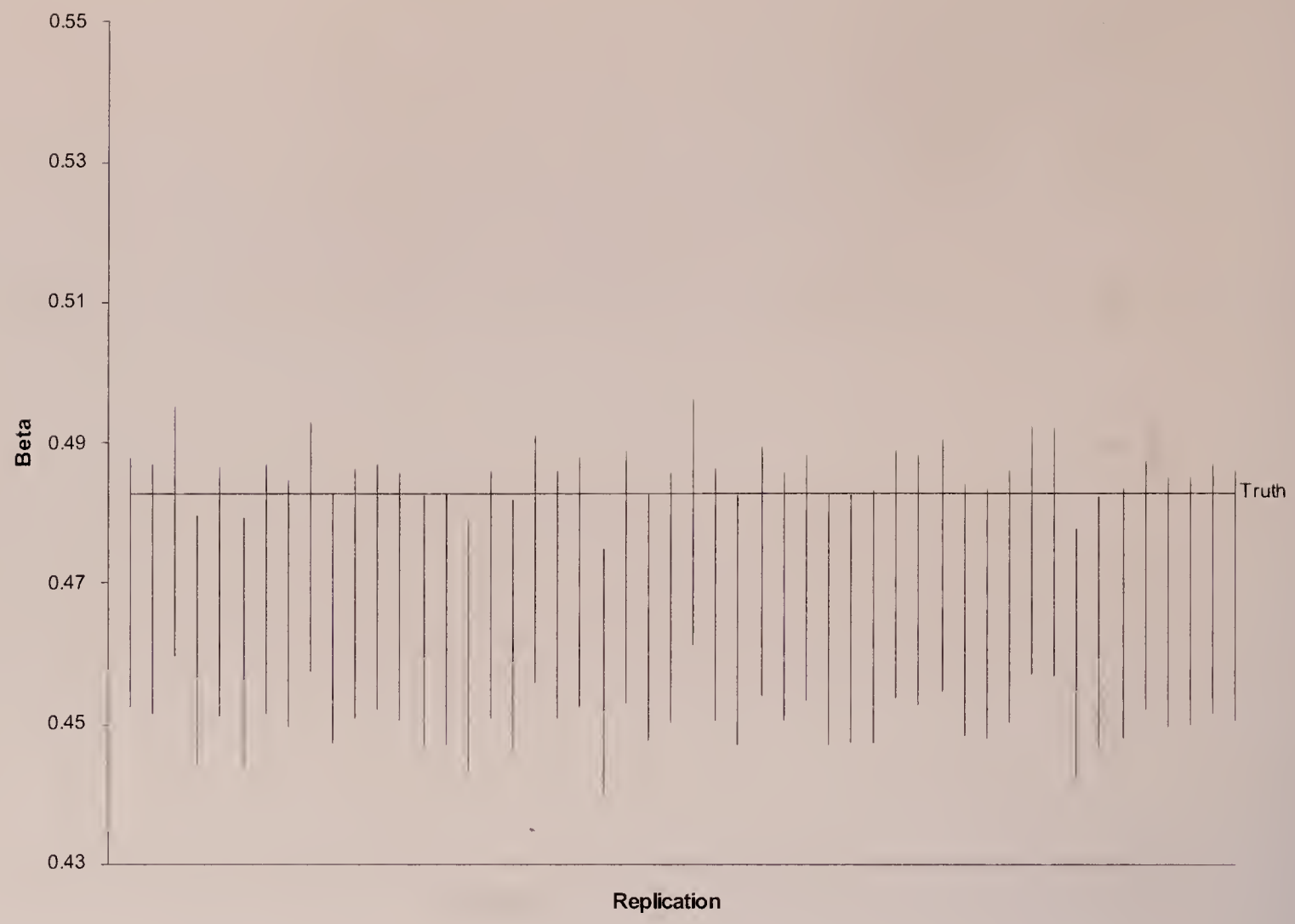

Figure 4.8. 80% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

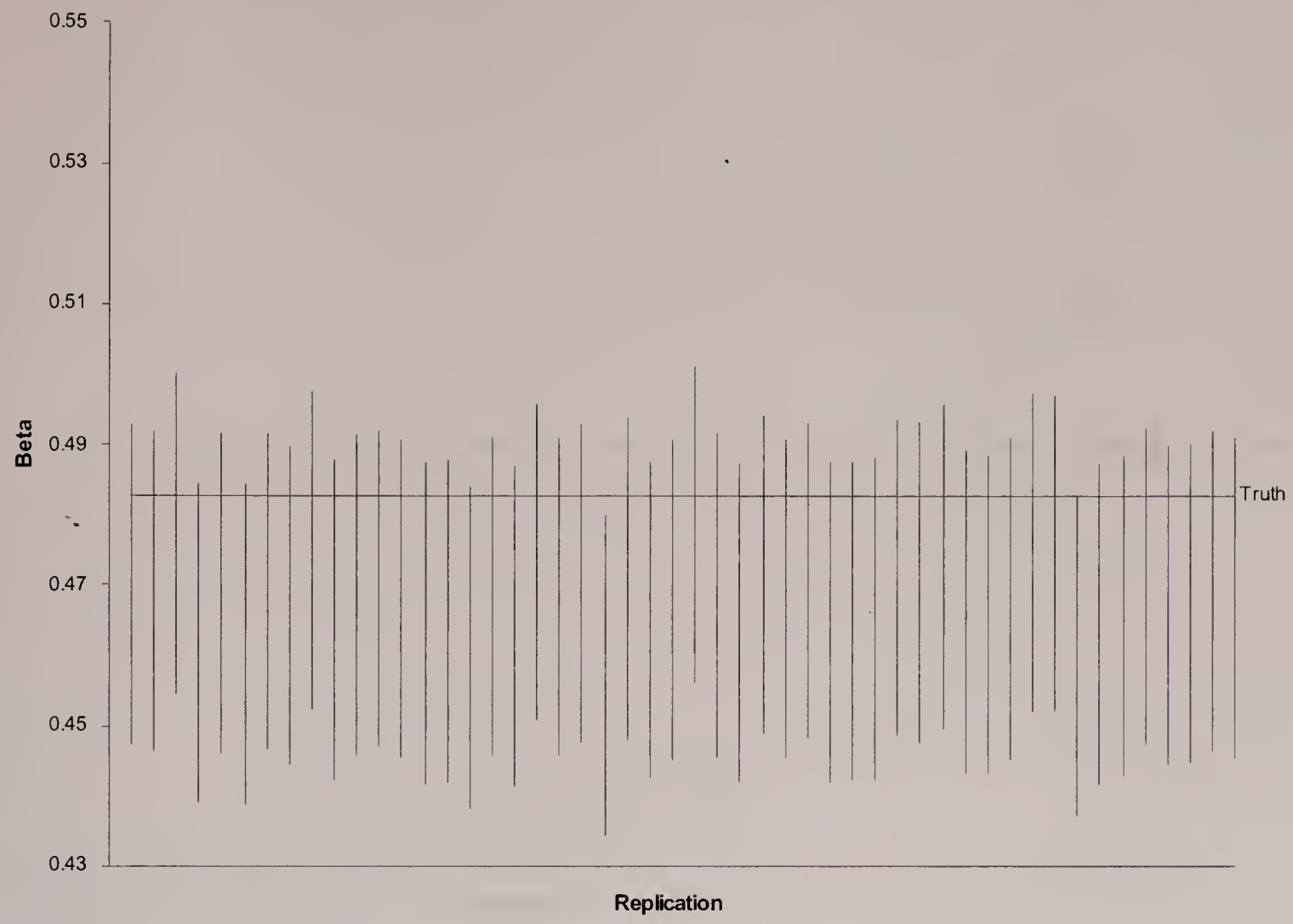

Figure 4.9. 90% Confidence Interval across 50 Replications for Condition 1 Using Post-hoc Approach

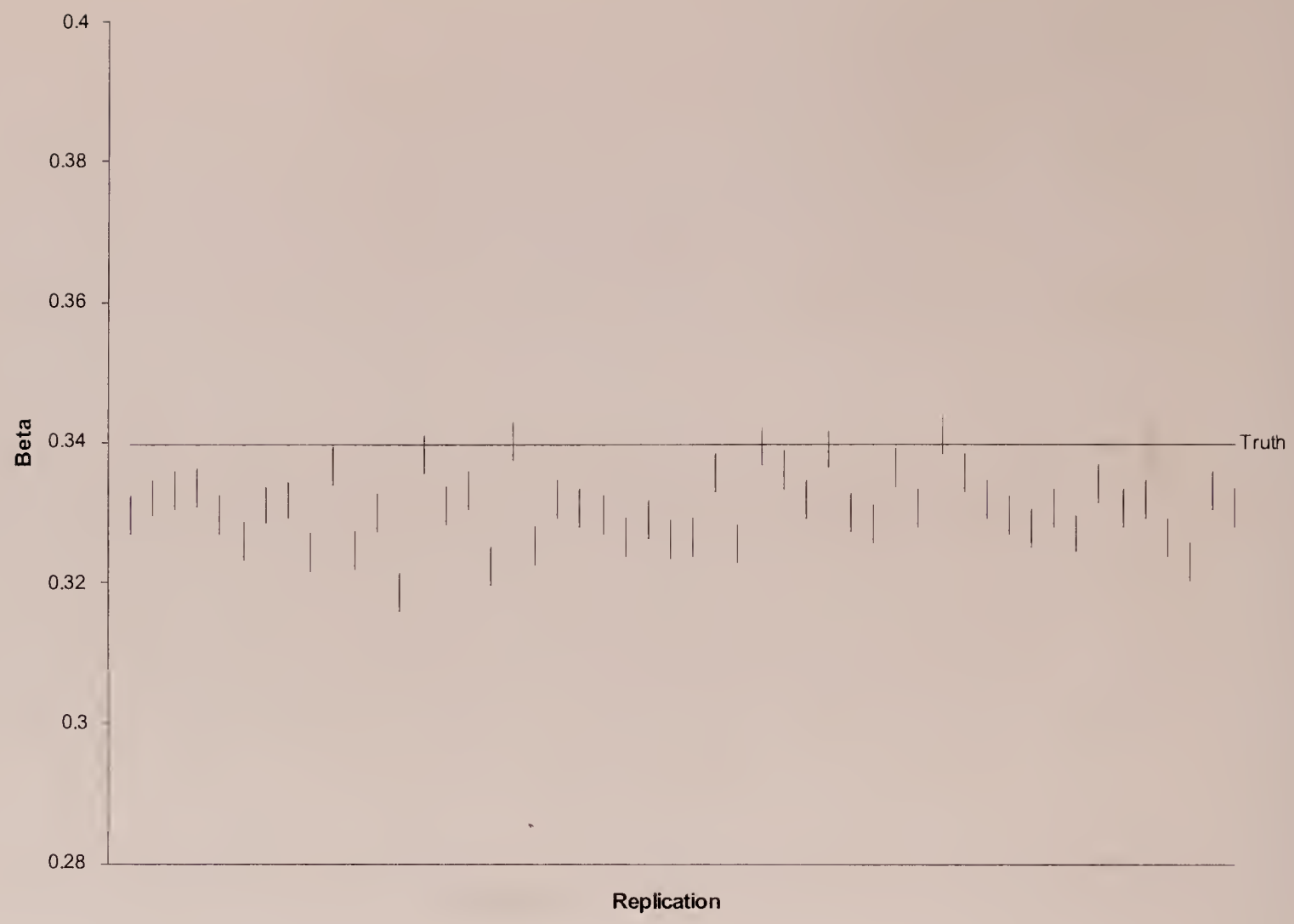

Figure 4.10. 10% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

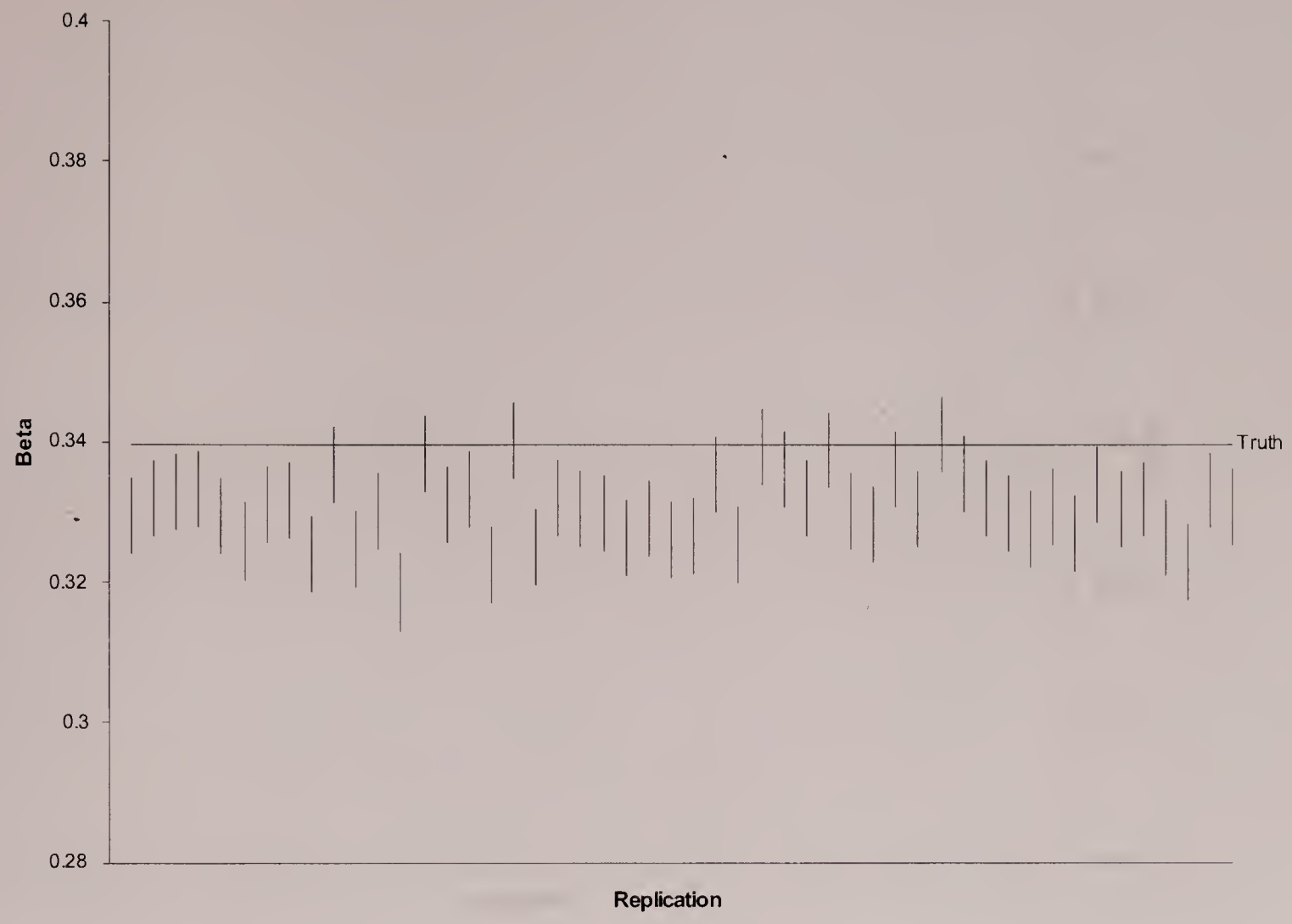

Figure 4.11. 20% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

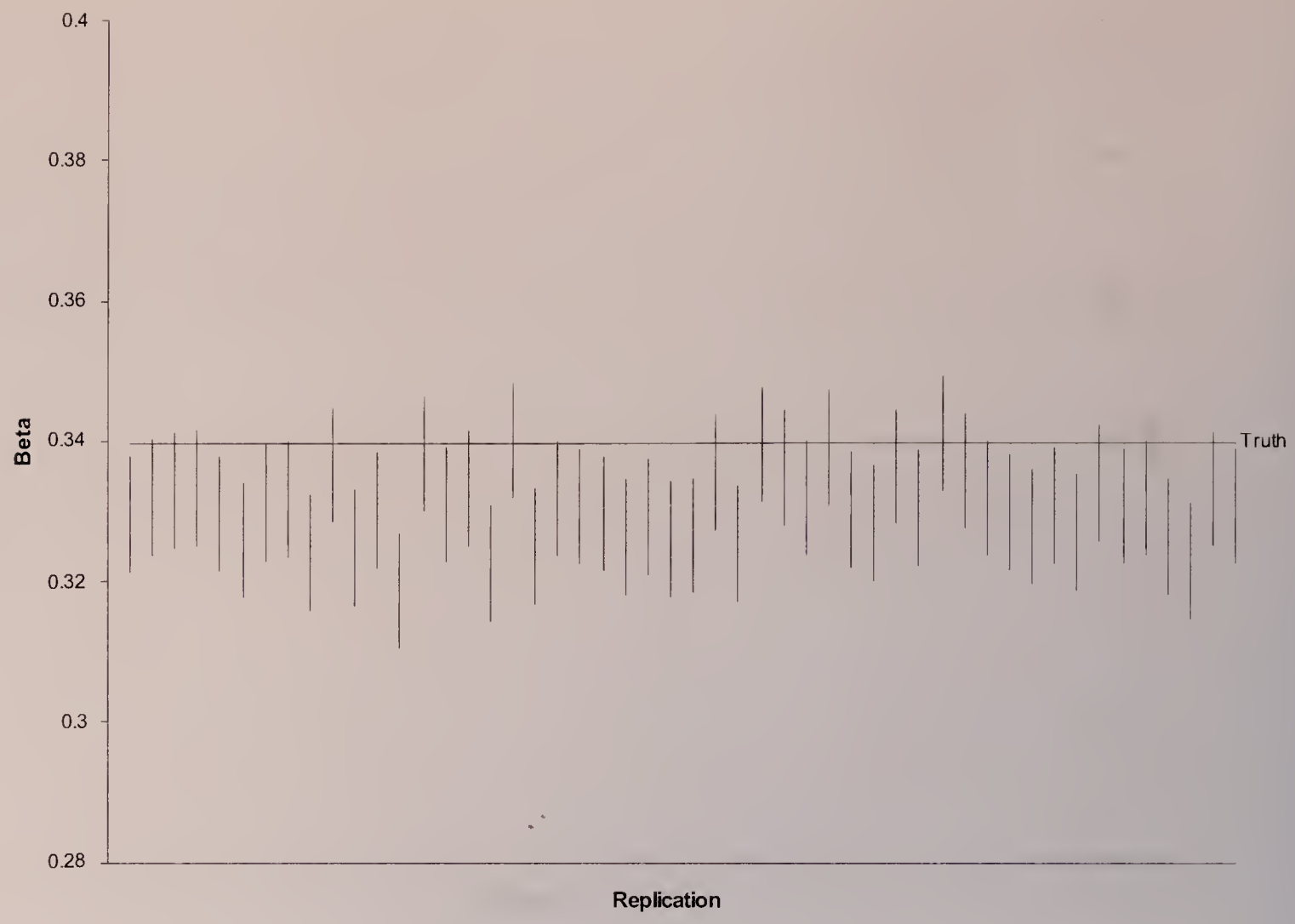

Figure 4.12. 30% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

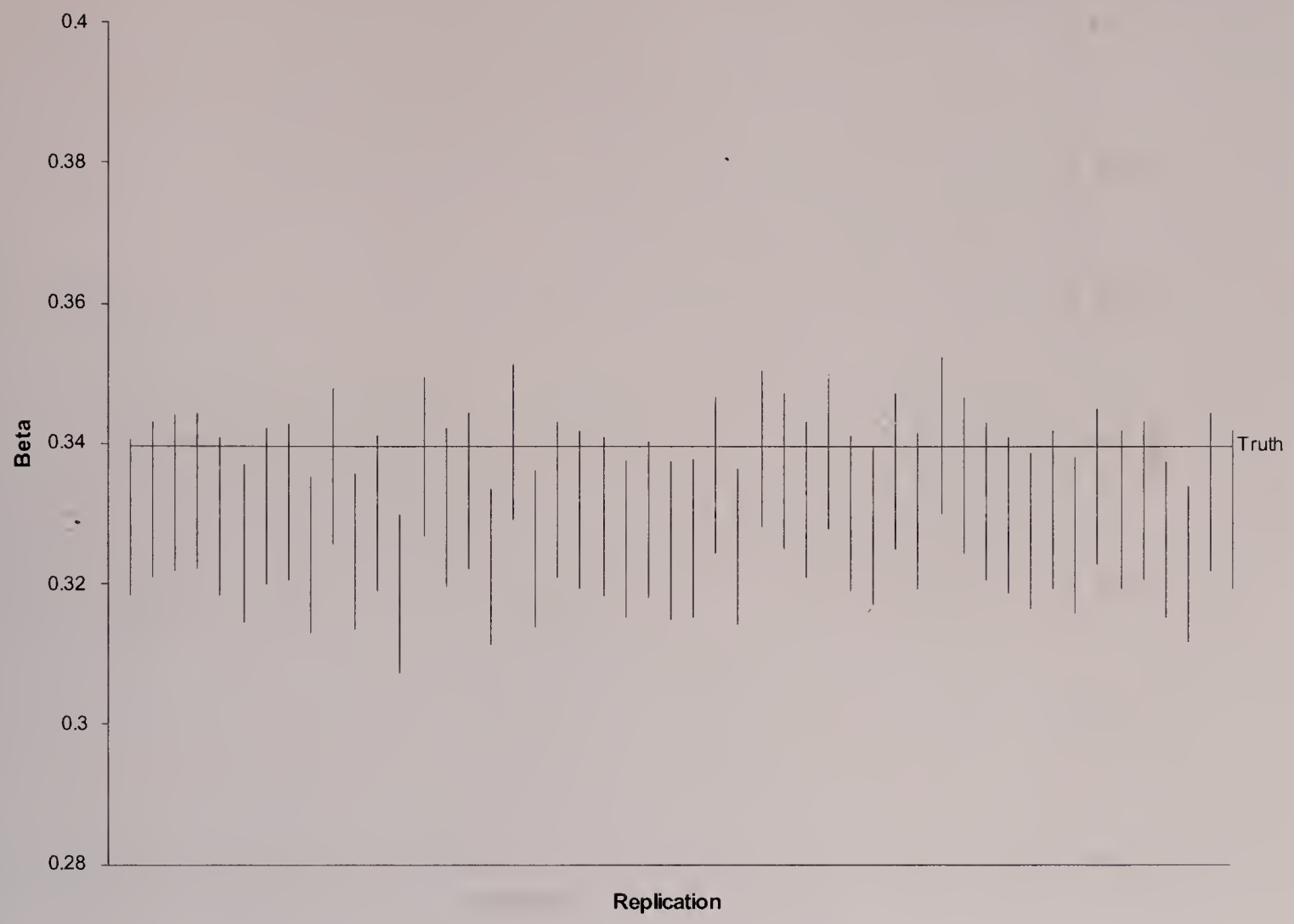

Figure 4.13. 40% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

Figure 4.14. 50% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

Figure 4.15. 60% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

Figure 4.16. 70% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

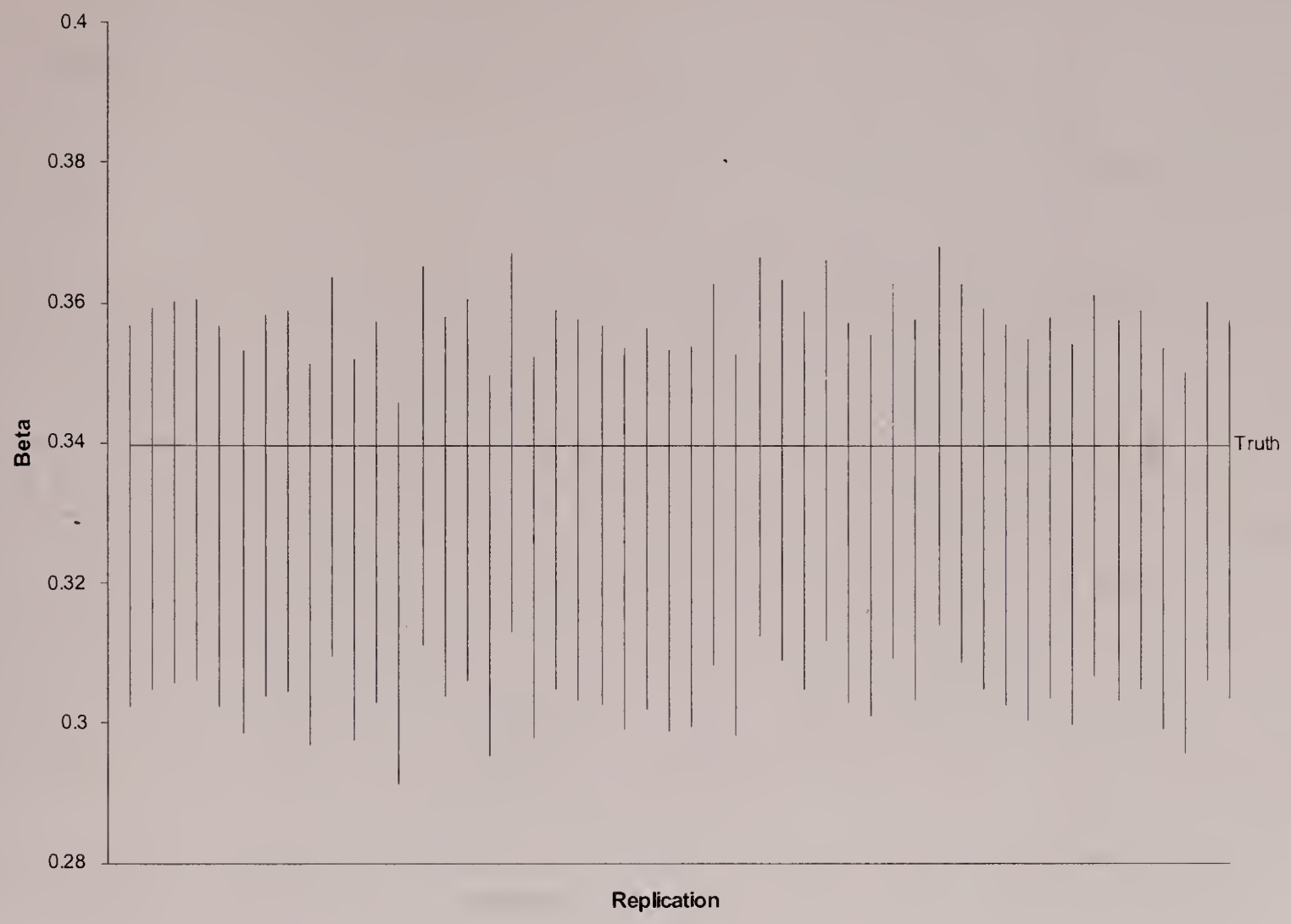

Figure 4.17. 80% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

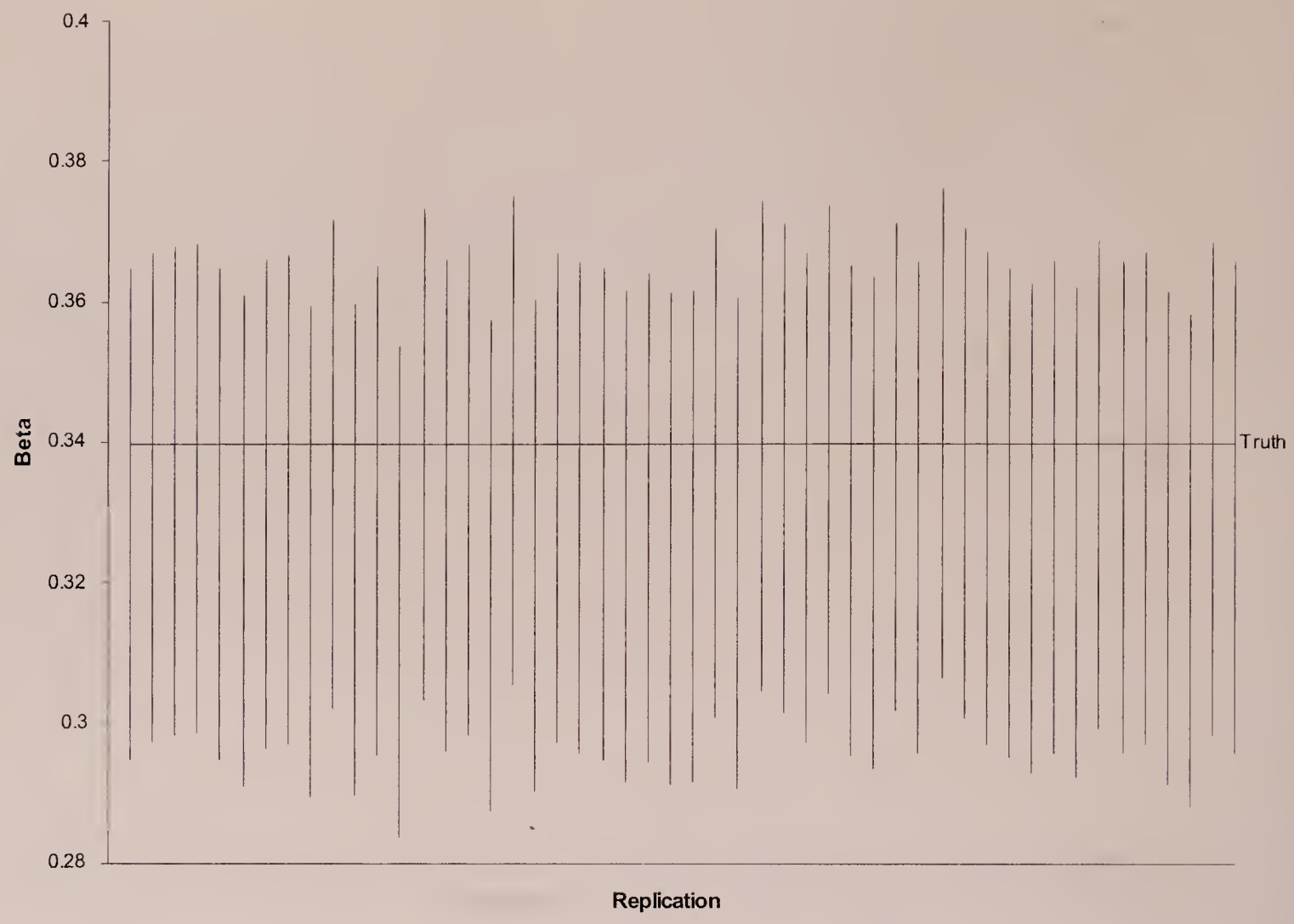

Figure 4.18. 90% Confidence Interval across 50 Replications for Condition 2 Using Post-hoc Approach

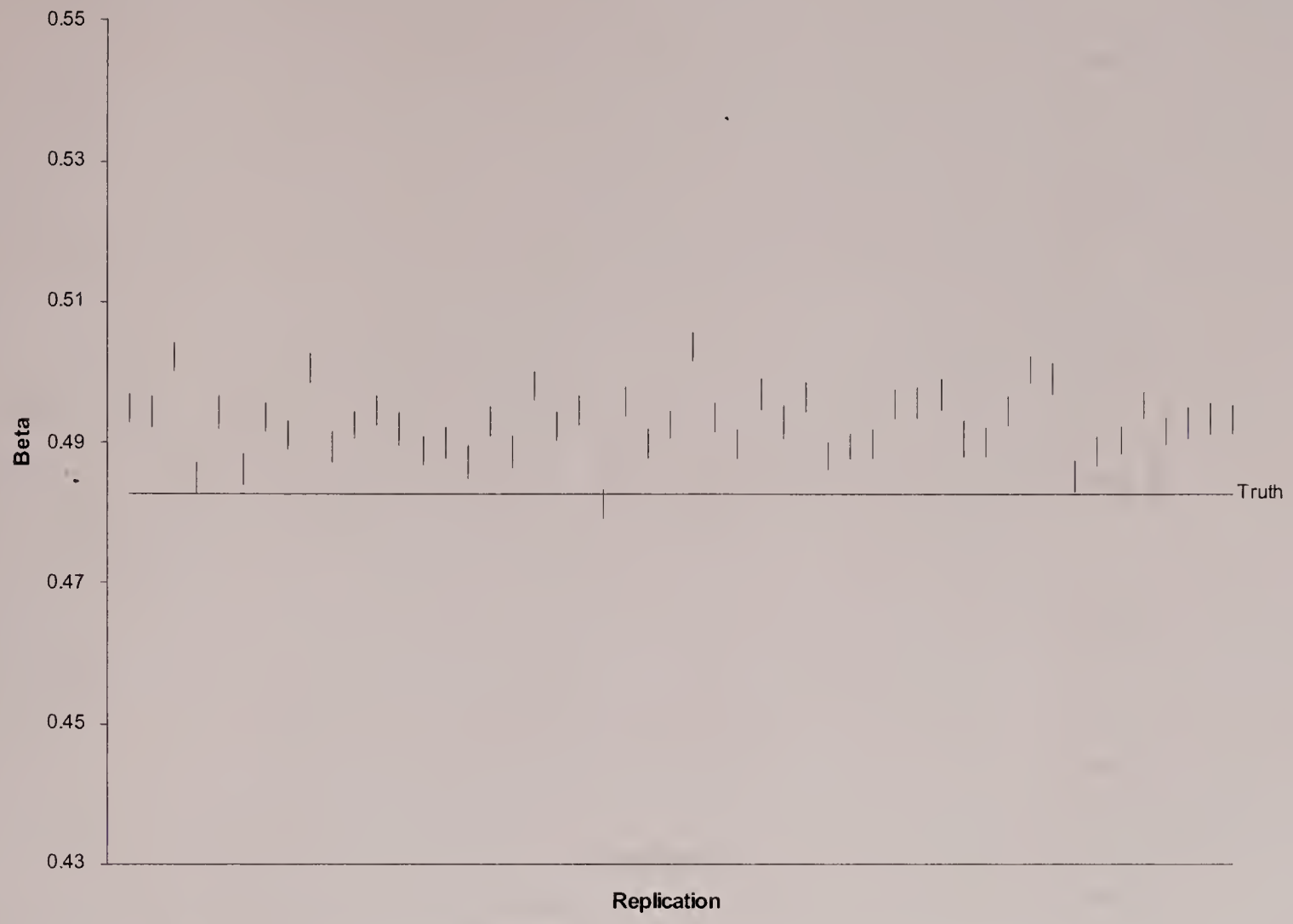

Figure 4.19. 10% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

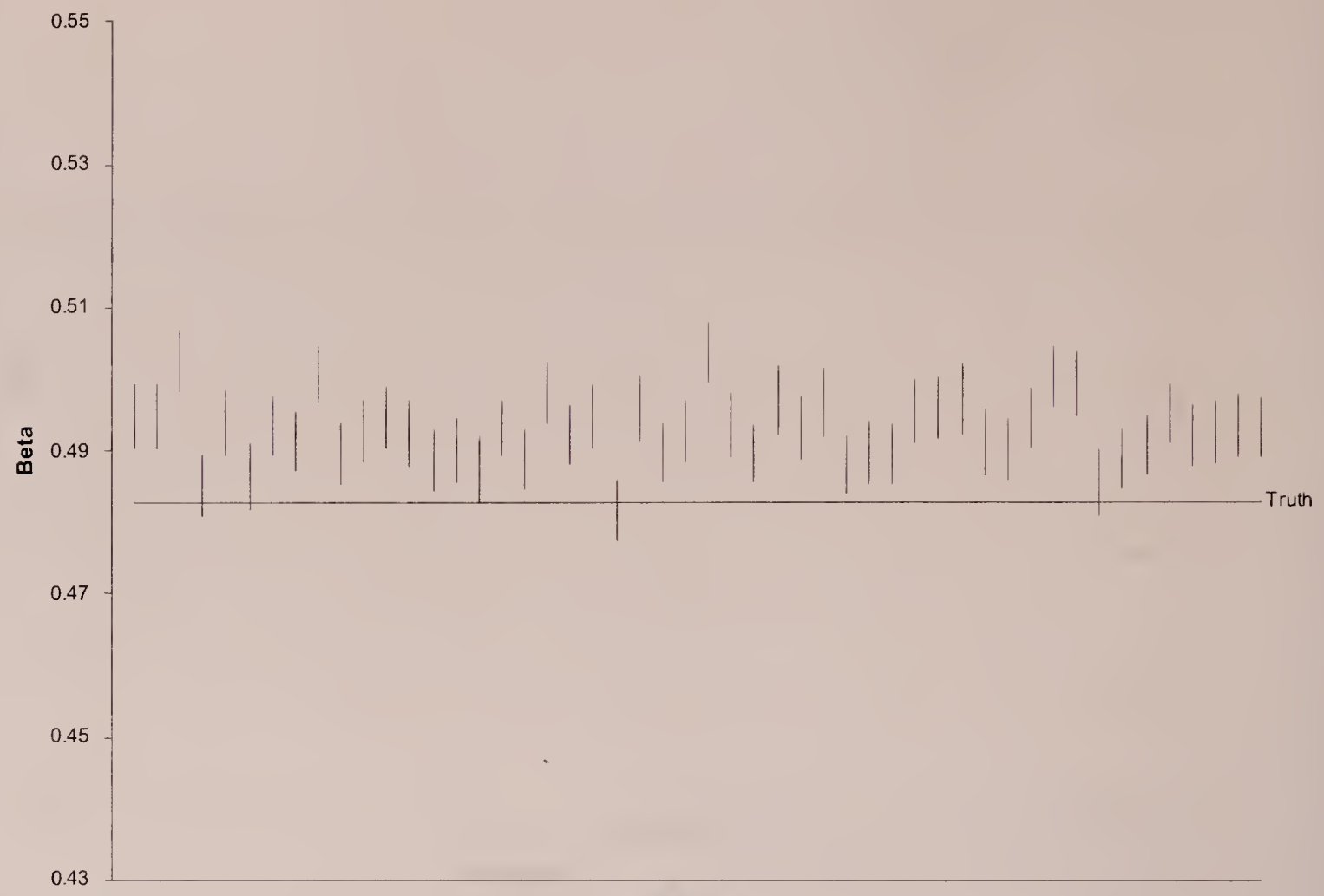

Figure 4.20. 20% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

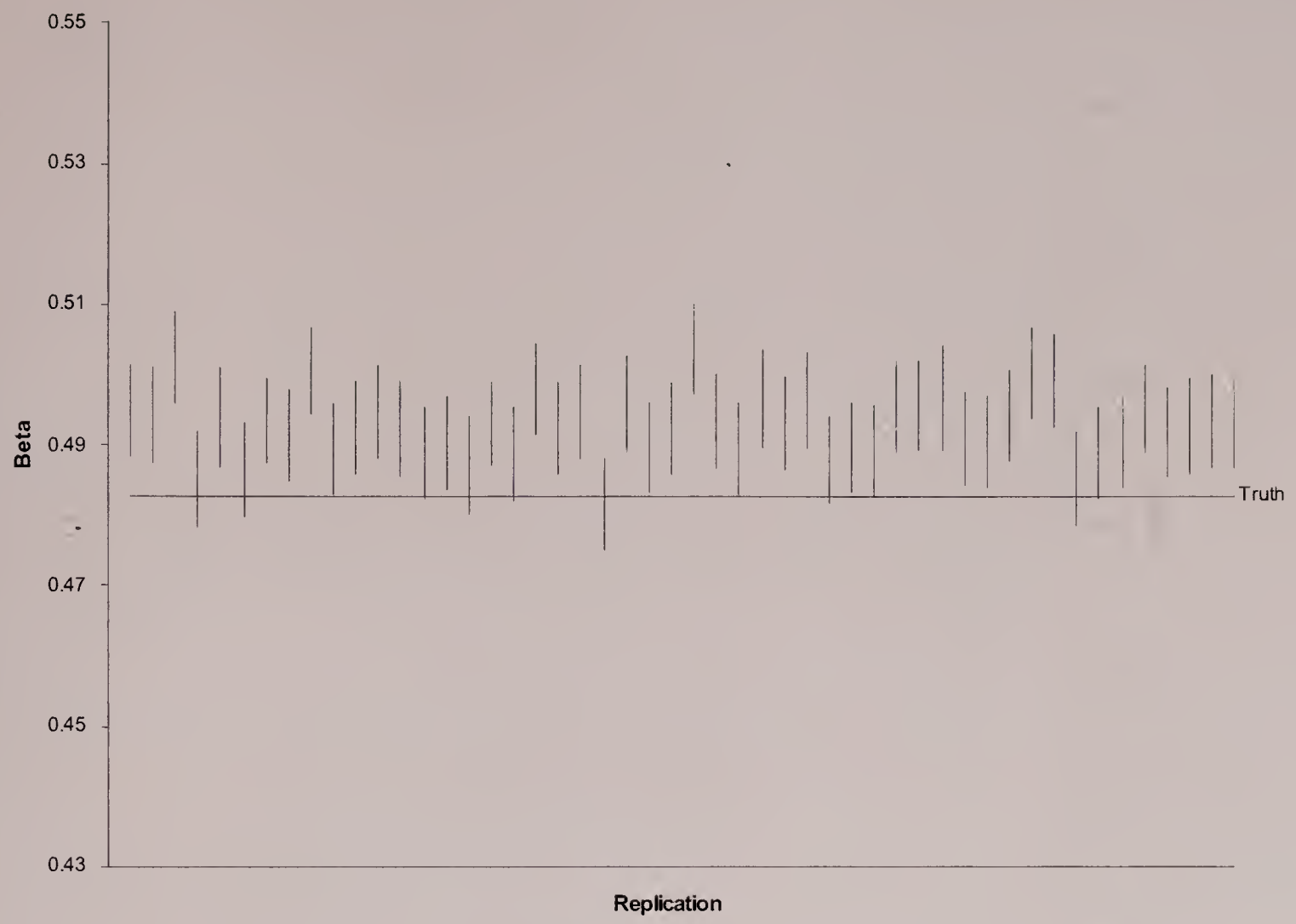

Figure 4.21. 30% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

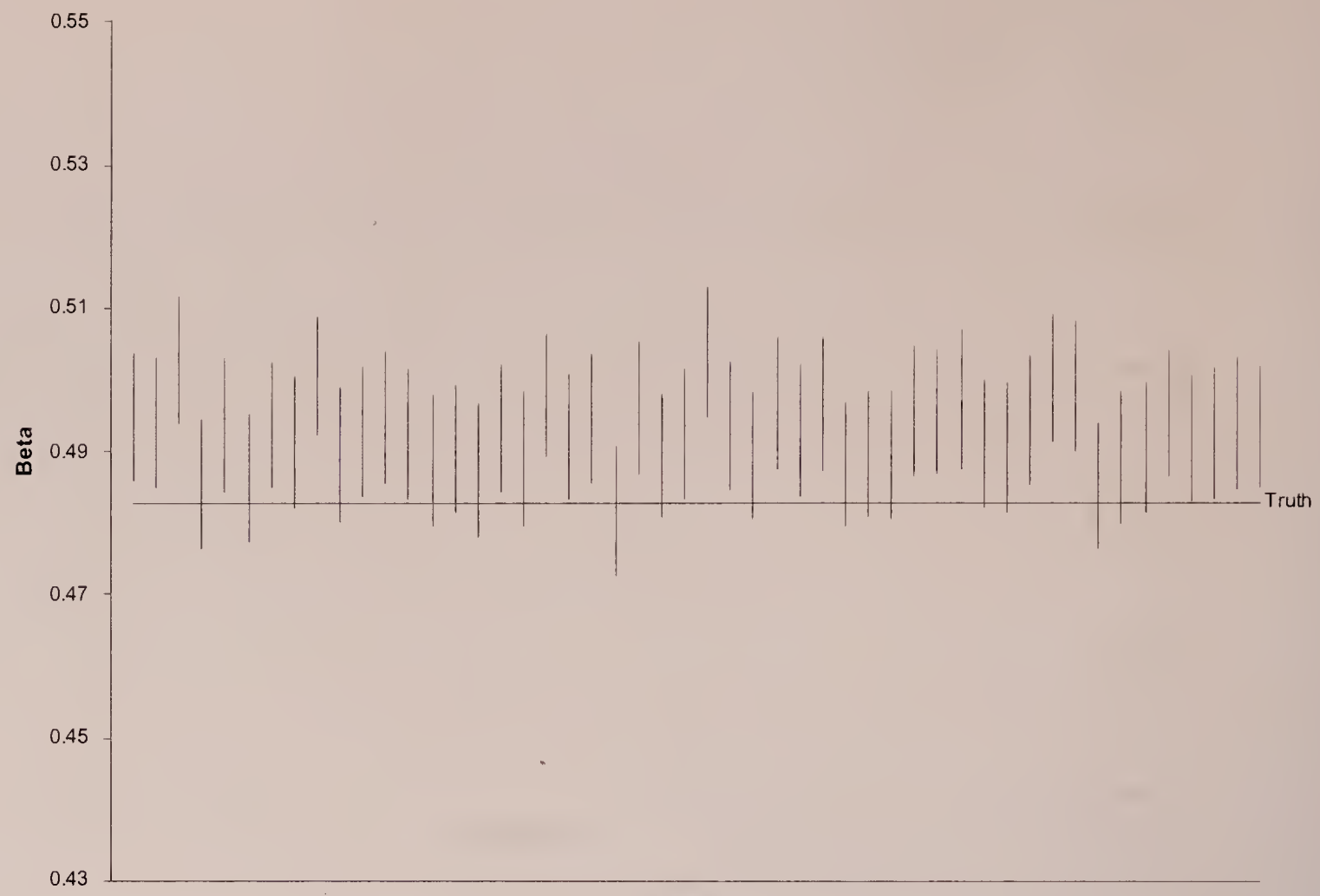

Figure 4.22. 40% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

Figure 4.23. 50% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

Figure 4.24. 60% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

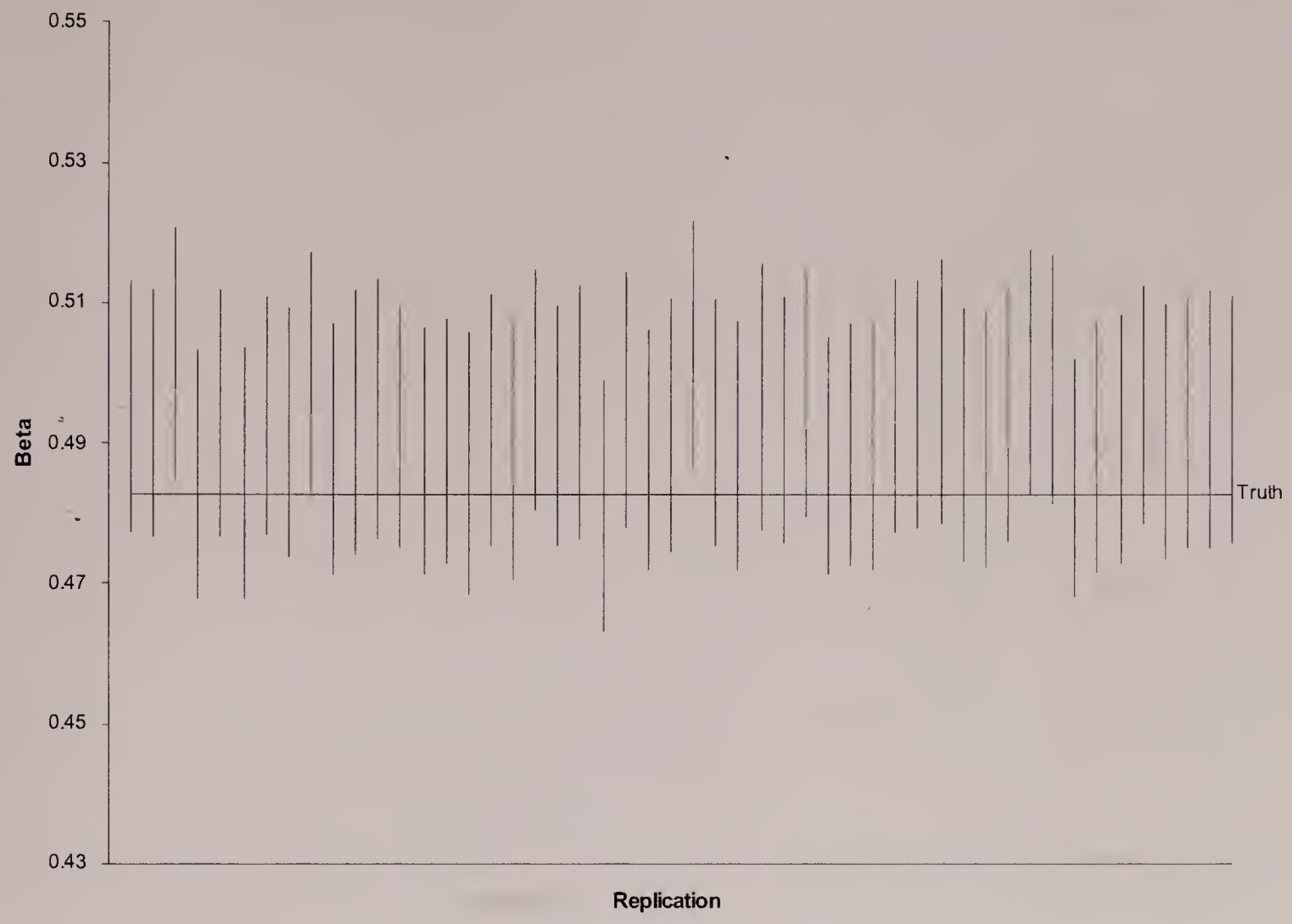

Figure 4.25. 70% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

Figure 4.26. 80 % Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

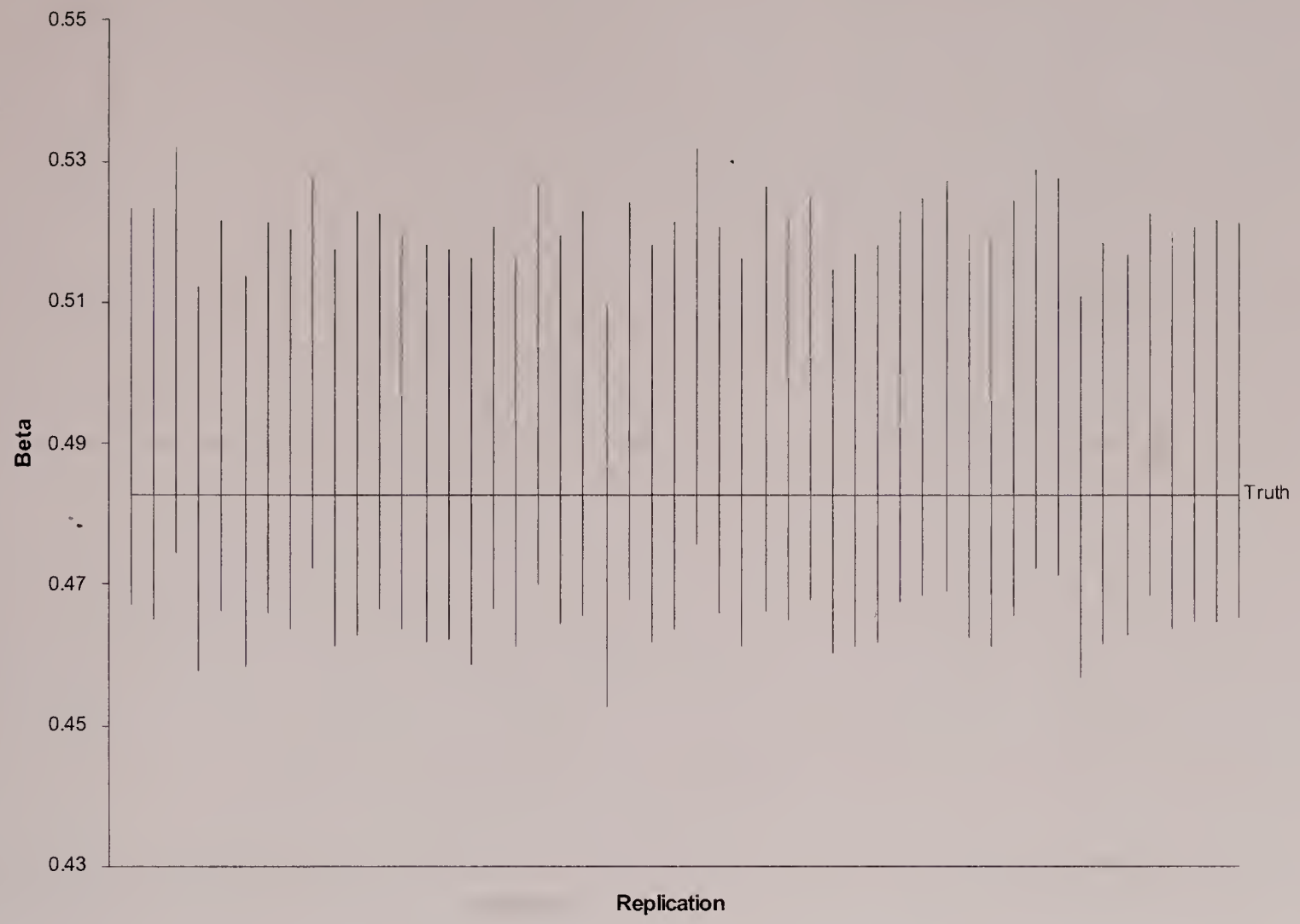

Figure 4.27. 90% Credible Interval across 50 Replications for Condition 1 Using Bayesian Approach

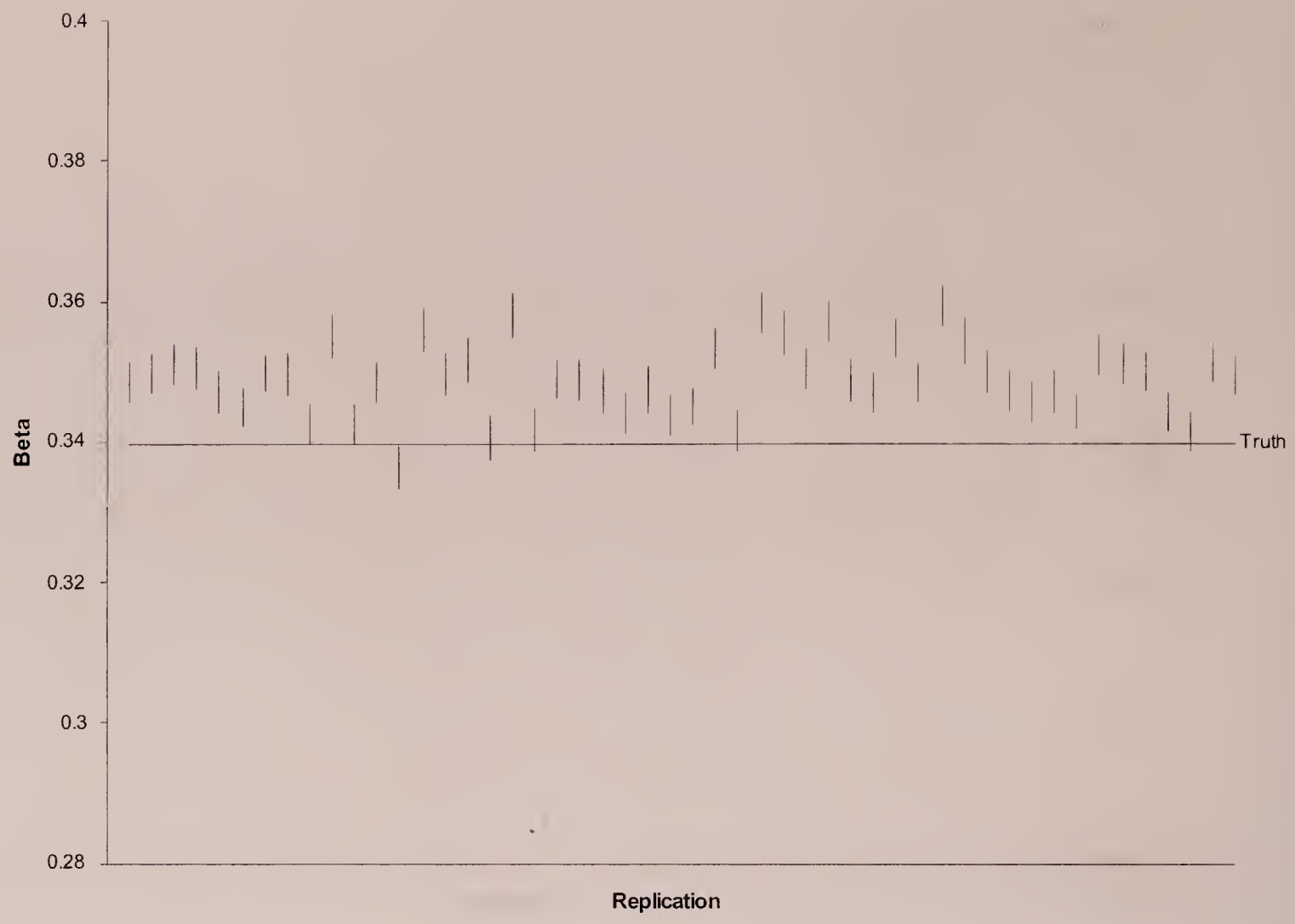

Figure 4.28. 10% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

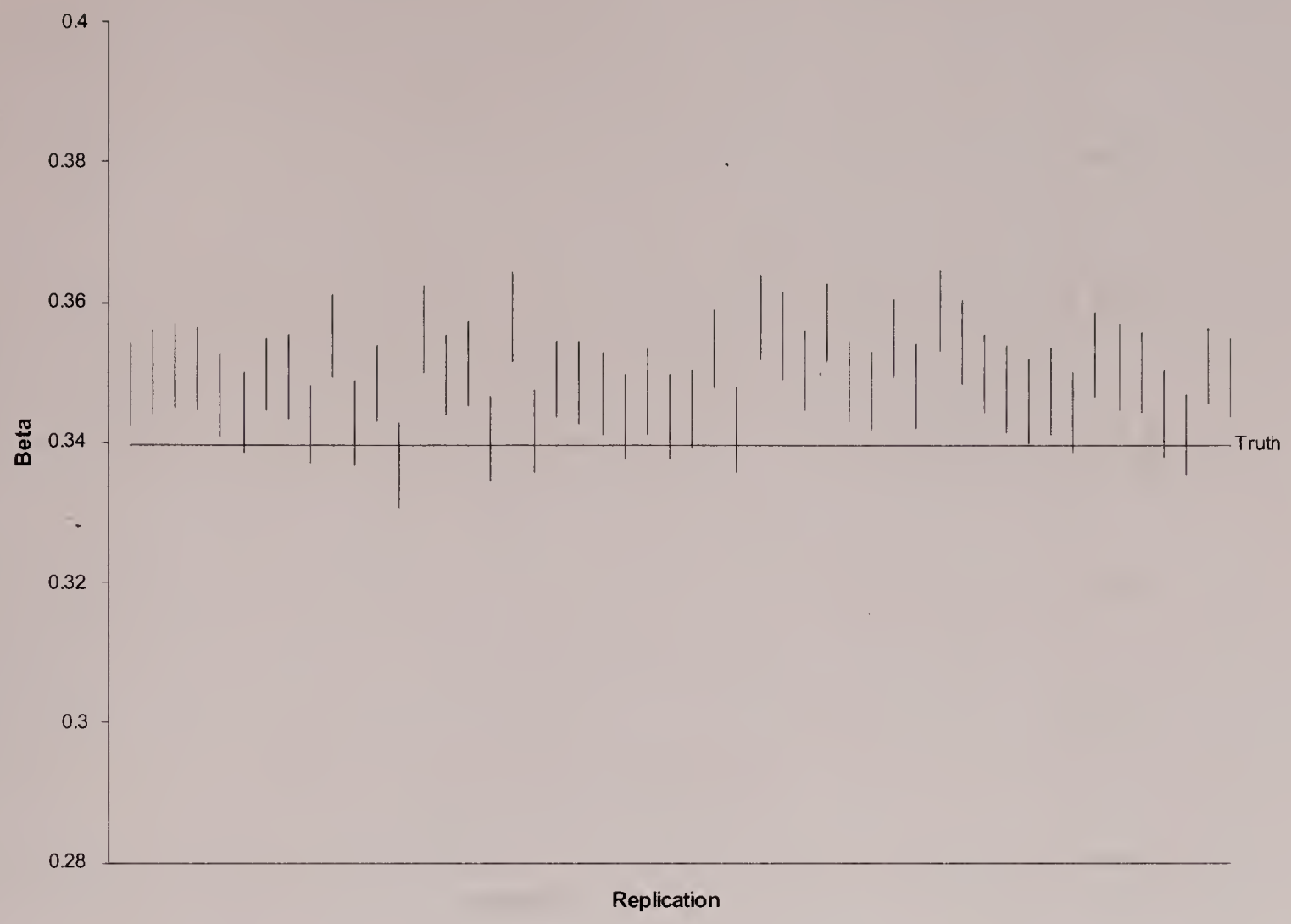

Figure 4.29. 20% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

Figure 4.30. 30% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

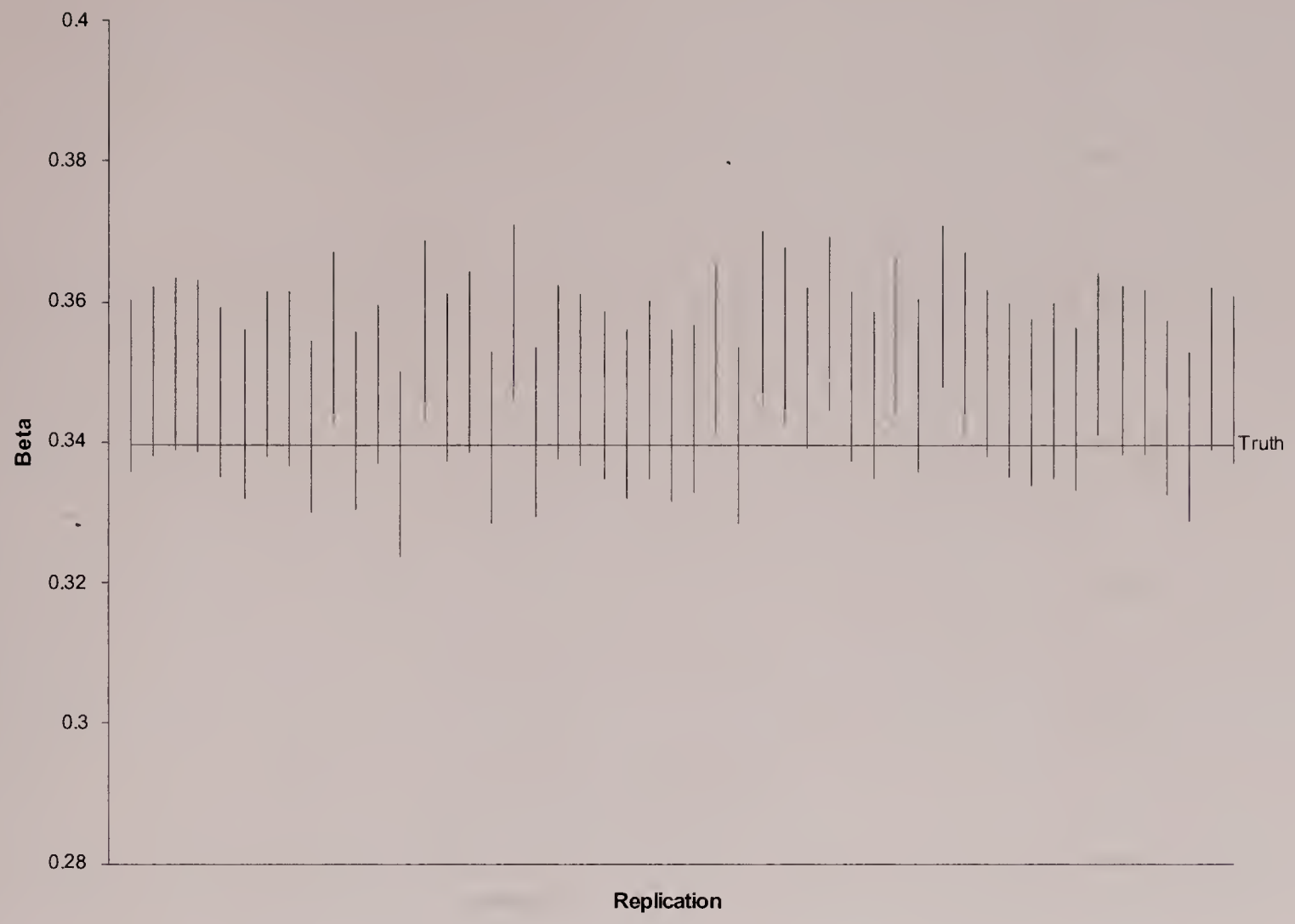

Figure 4.31. 40% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

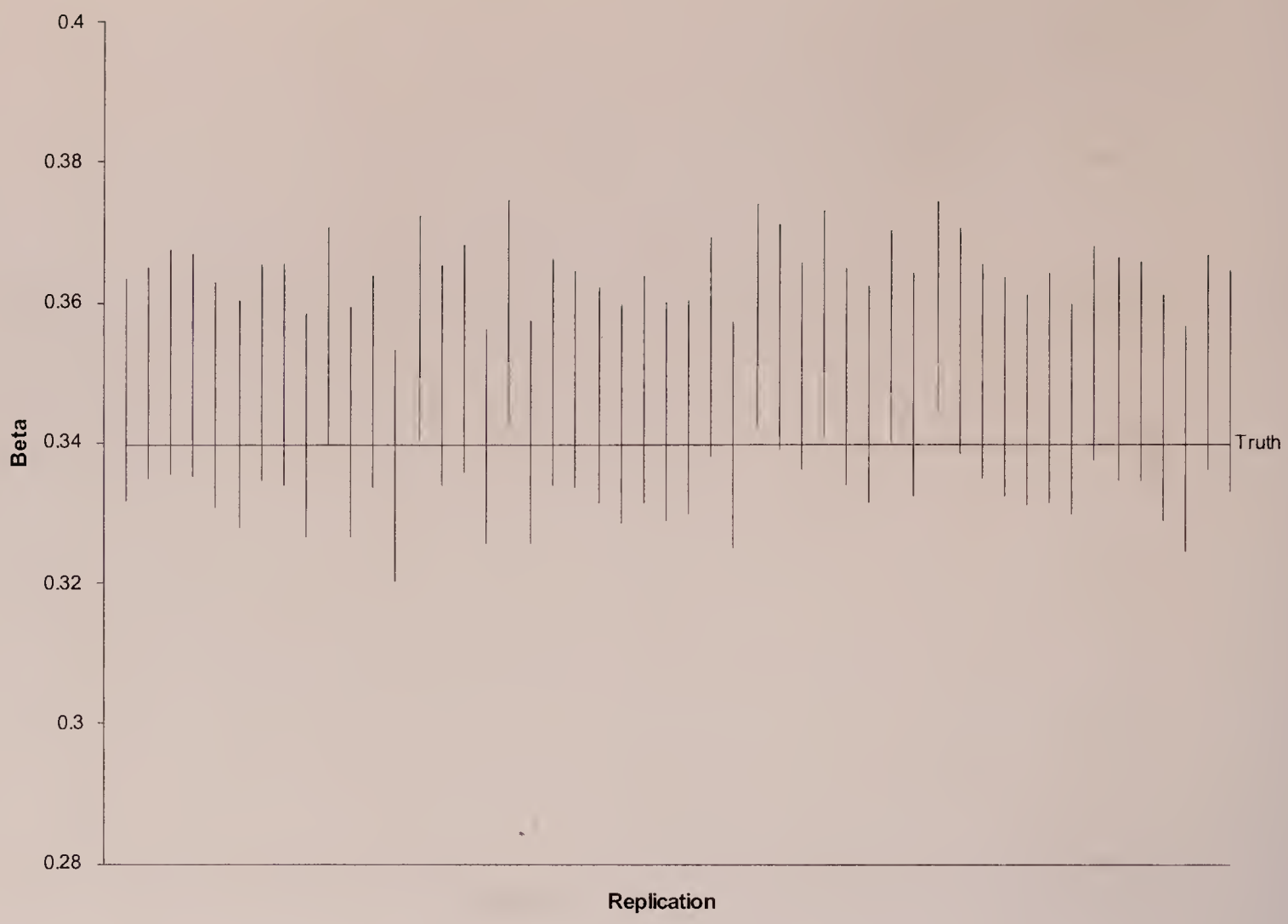

Figure 4.32. 50% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

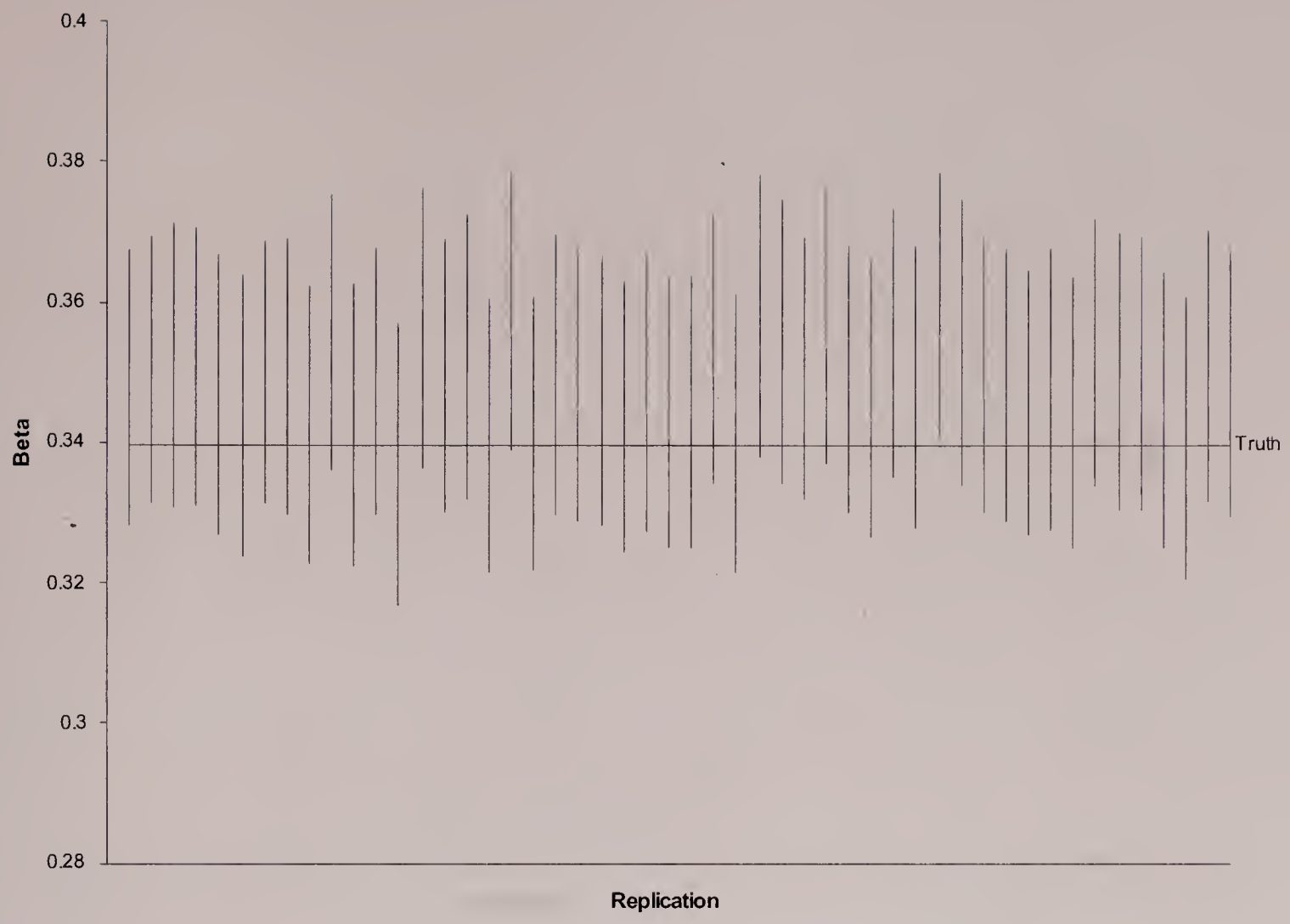

Figure 4.33. 60% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

Figure 4.34. 70% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

Figure 4.35. 80% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

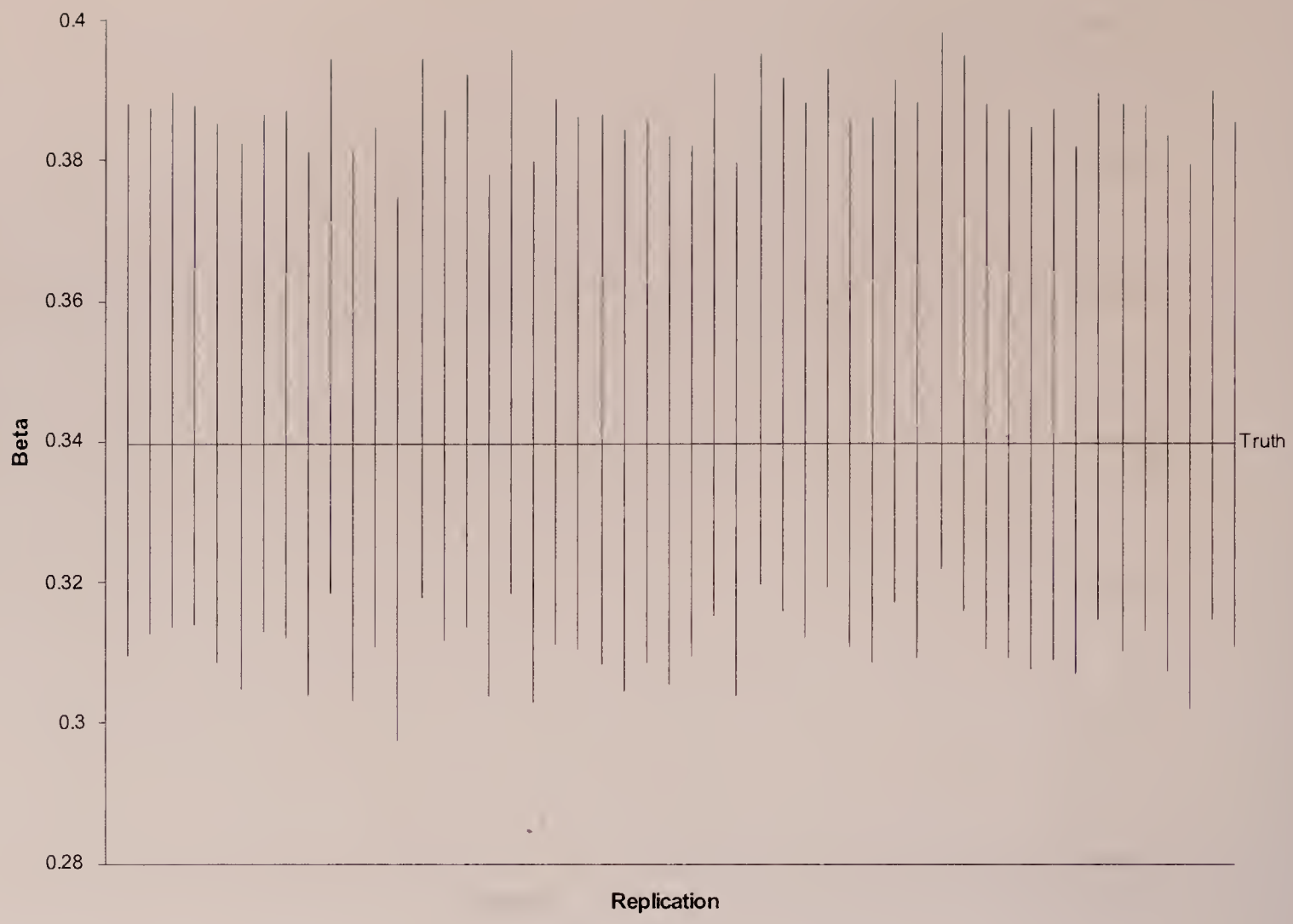

Figure 4.36. 90% Credible Interval across 50 Replications for Condition 2 Using Bayesian Approach

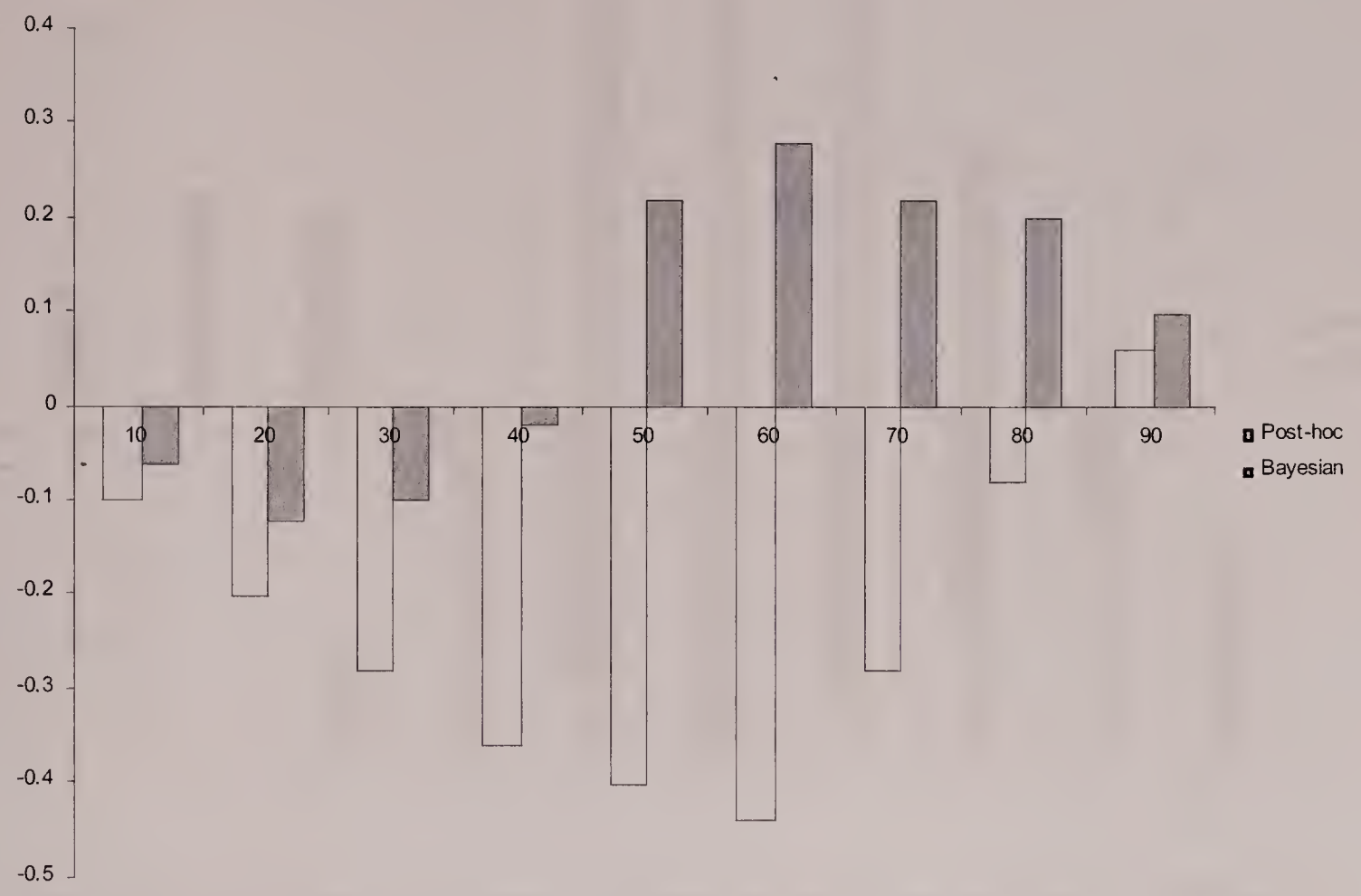

Figure 4.37. The Difference between the Observed and Expected Coverage Probability at Each Interval for Condition 1 Covariate 1

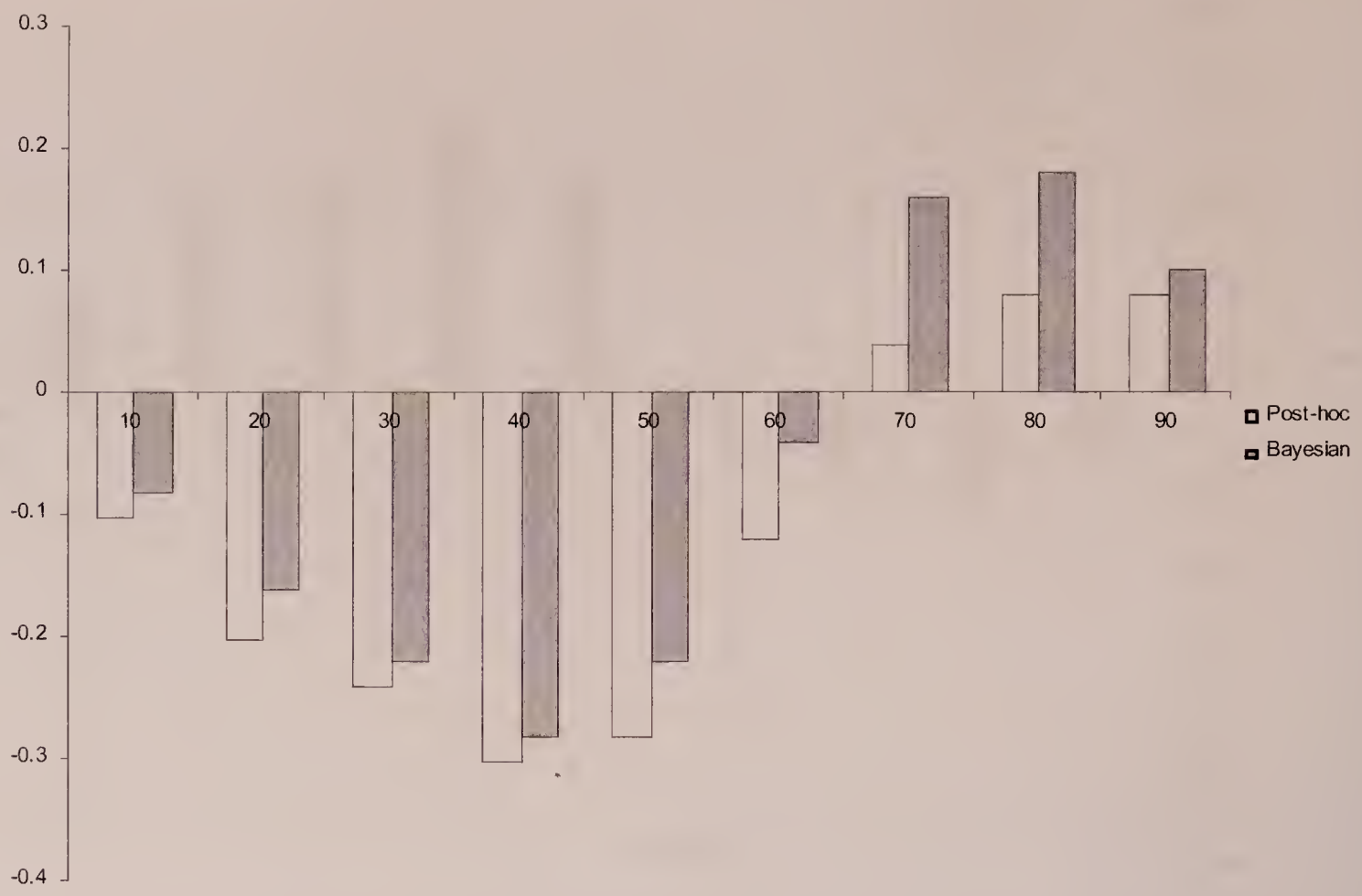

Figure 4.38. The Difference between the Observed and Expected Coverage Probability at Each Interval for Condition 1 Covariate 2

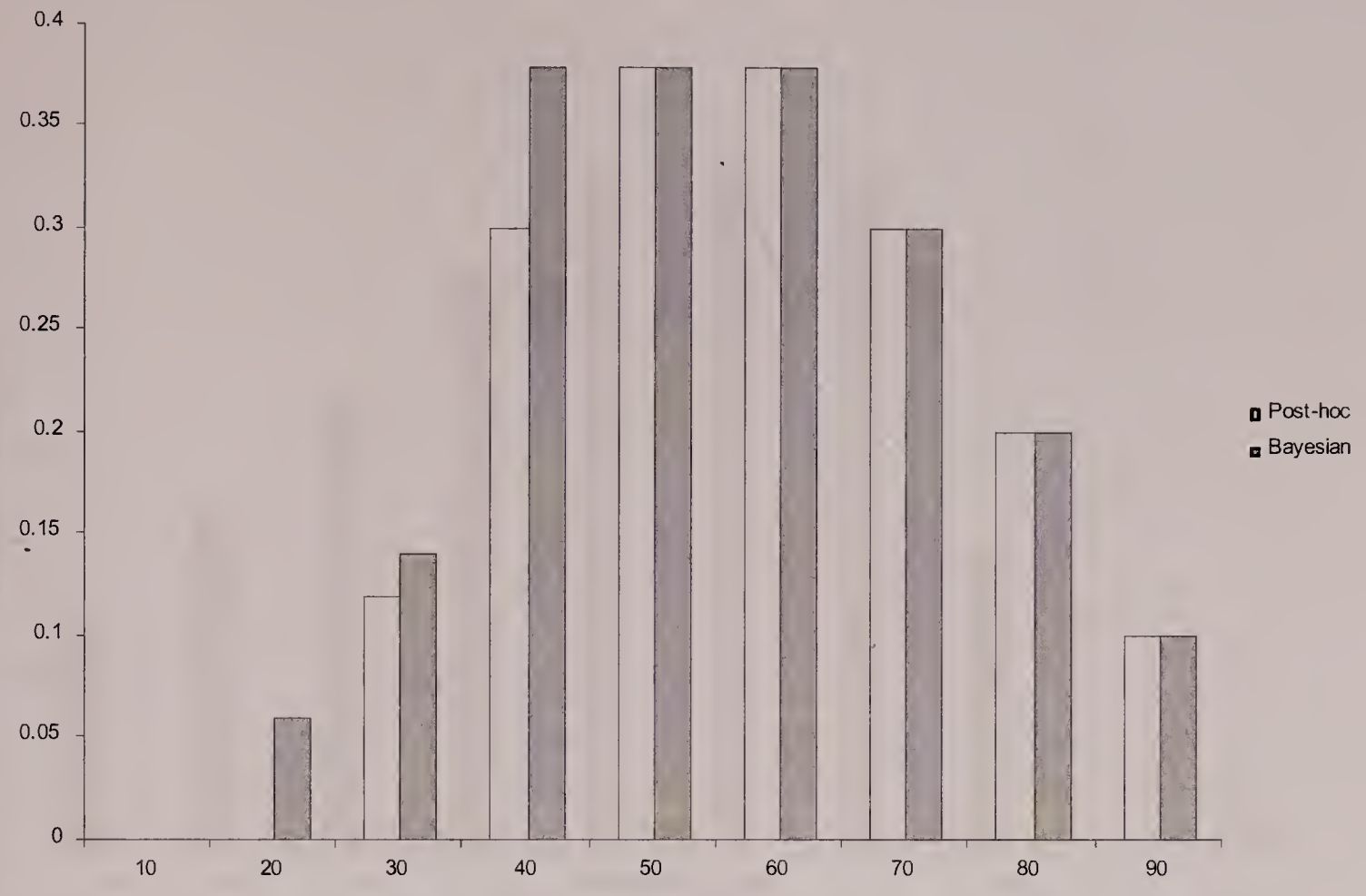

Figure 4.39. The Difference between the Observed and Expected Coverage Probability at Each Interval for Condition 2 Covariate 1

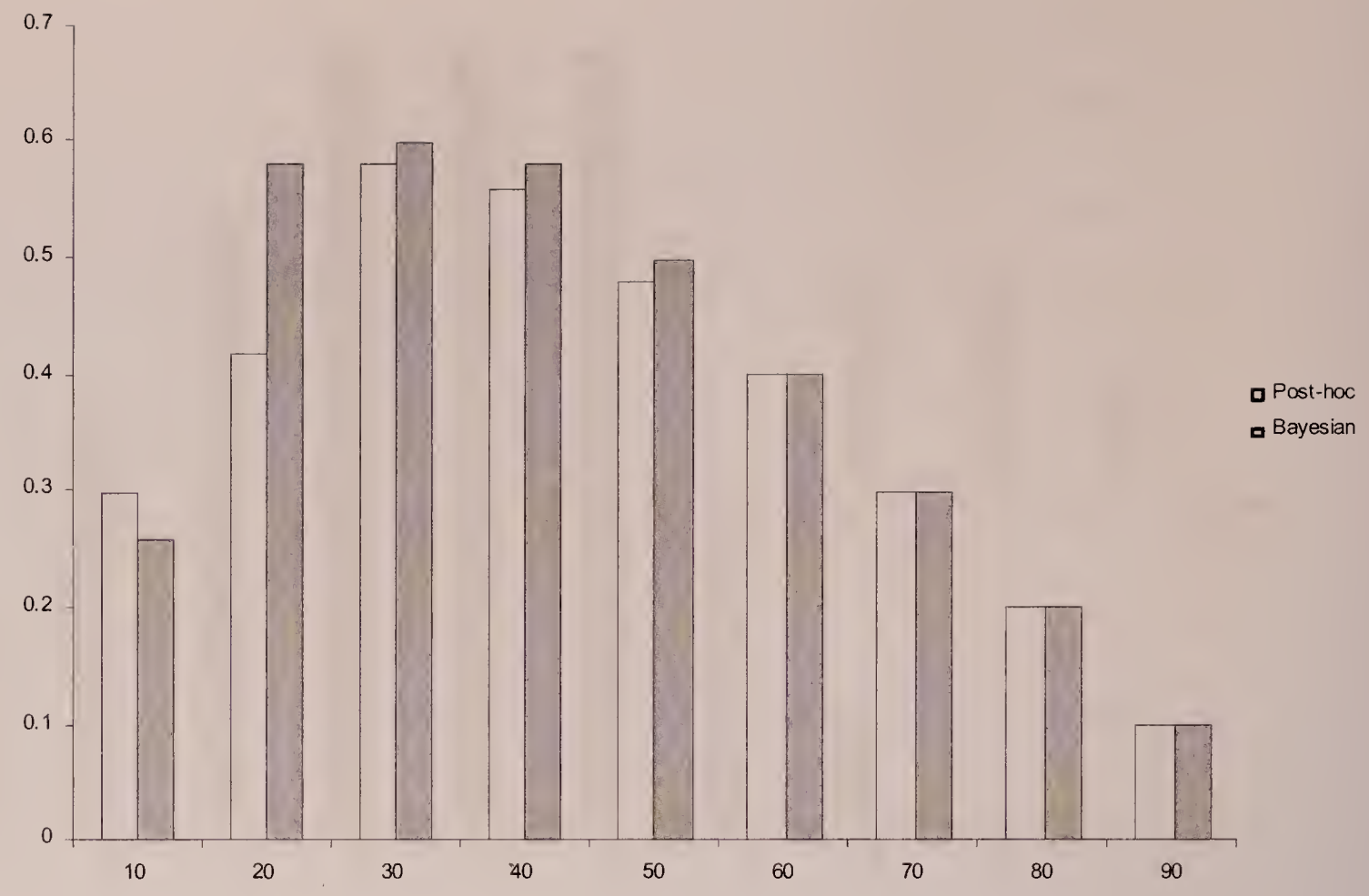

Figure 4.40. The Difference between the Observed and Expected Coverage Probability at Each Interval for Condition 2 Covariate 2

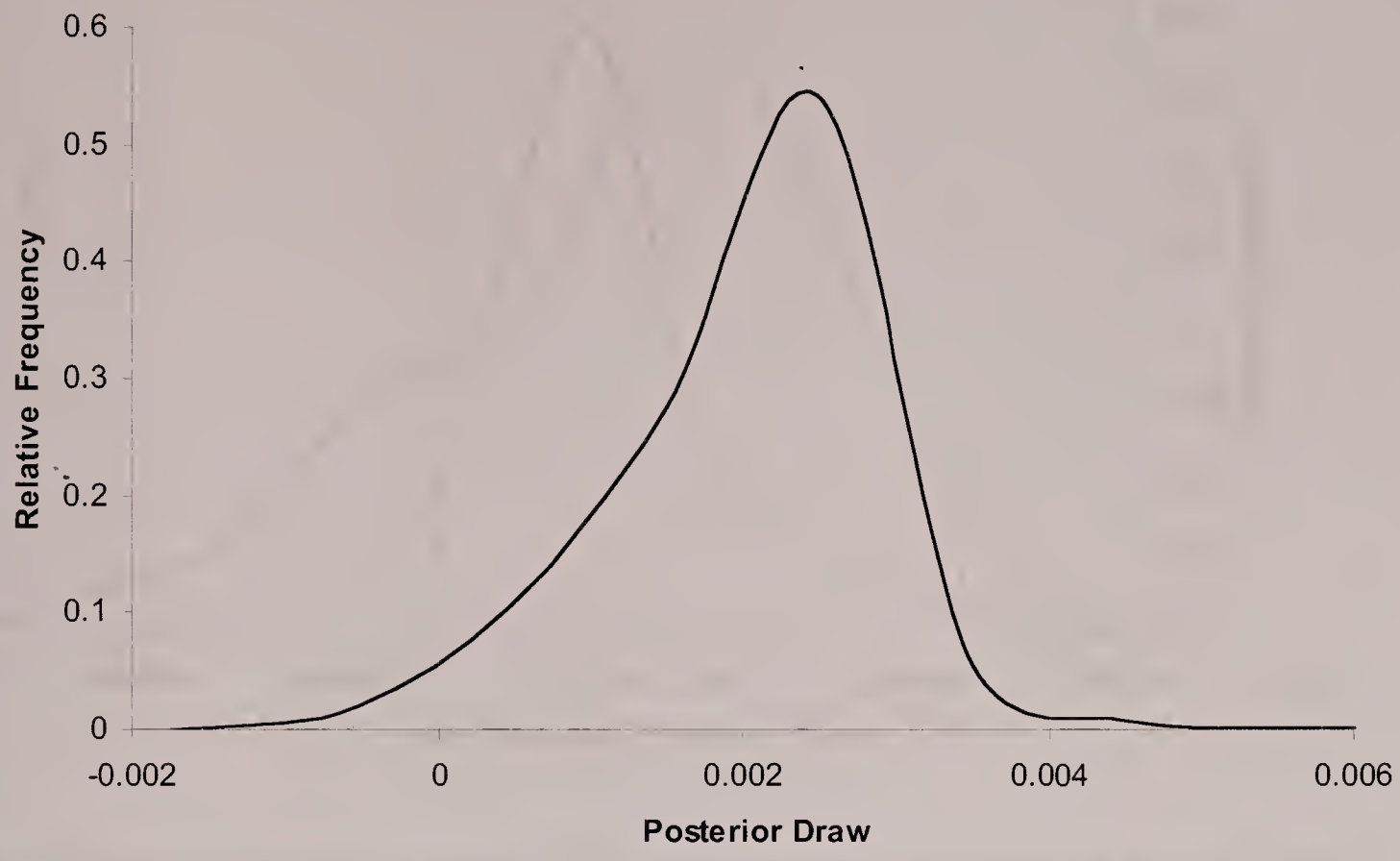

Figure 4.41. Posterior Distribution of Coefficient of Covariate Vignette Word Count for a-parameter

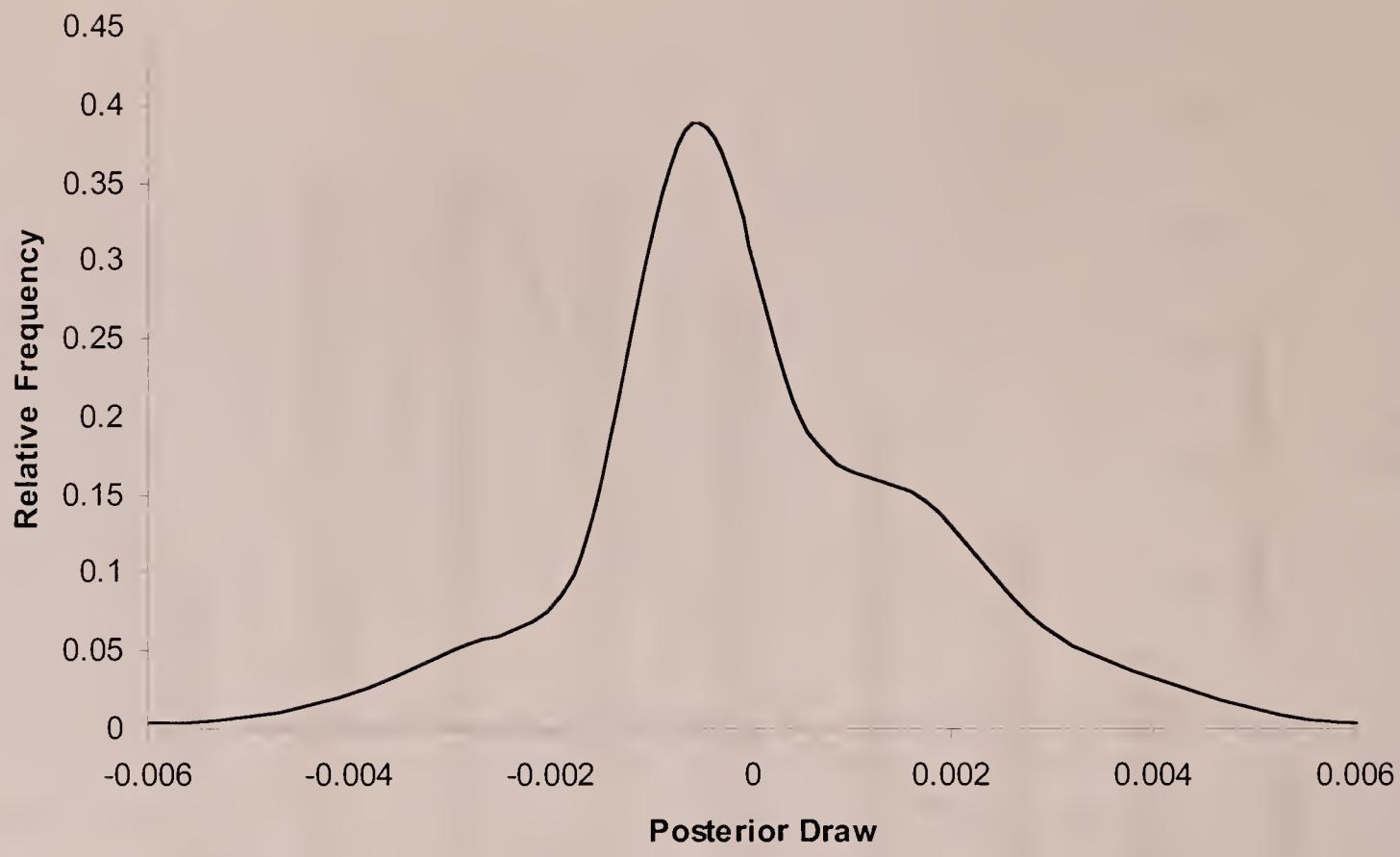

Figure 4.42. Posterior Distribution of Coefficient of Covariate Stem Word Count for a-parameter

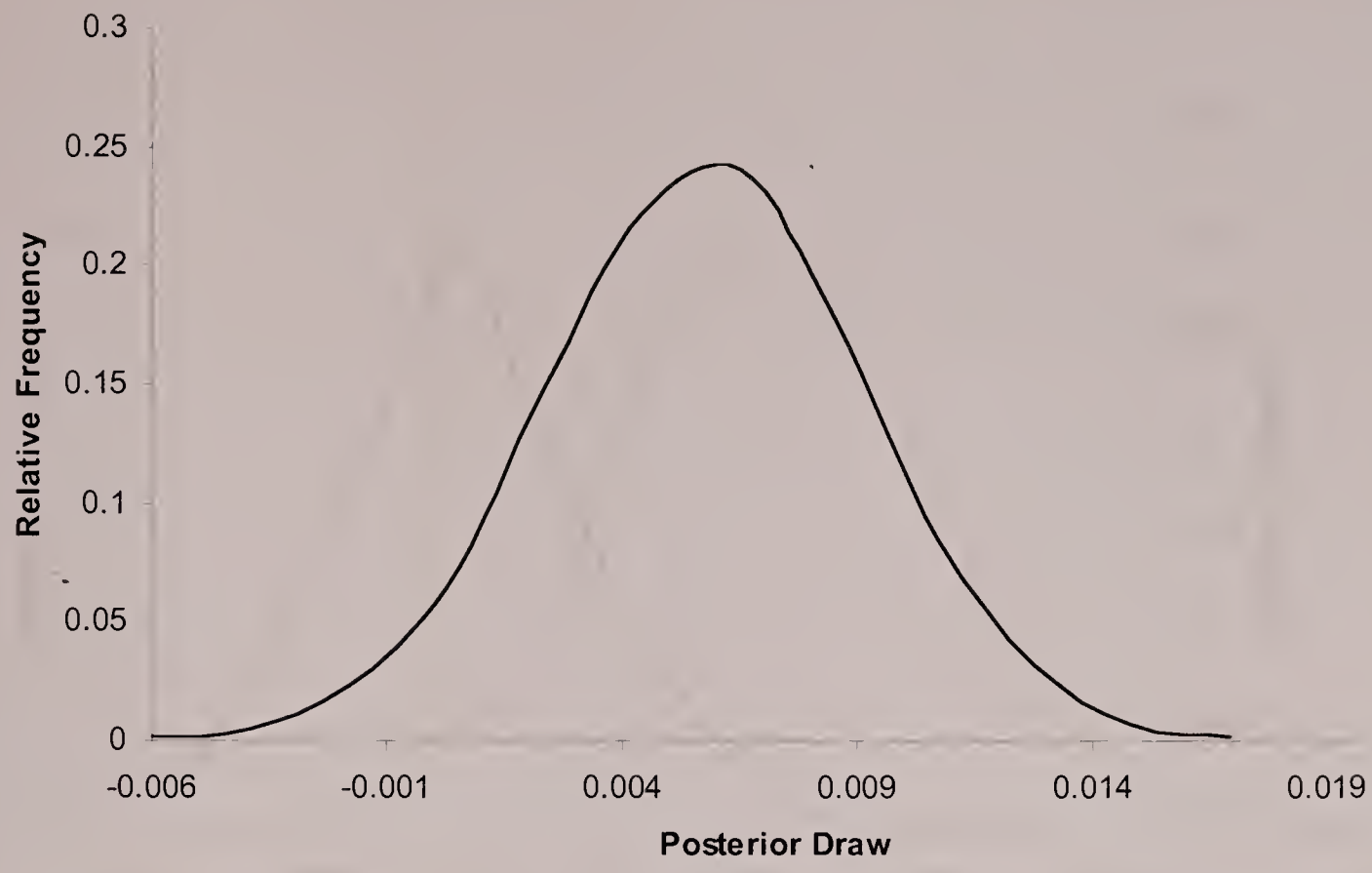

Figure 4.43. Posterior Distribution of Coefficient of Covariate Option Word Count for a-parameter

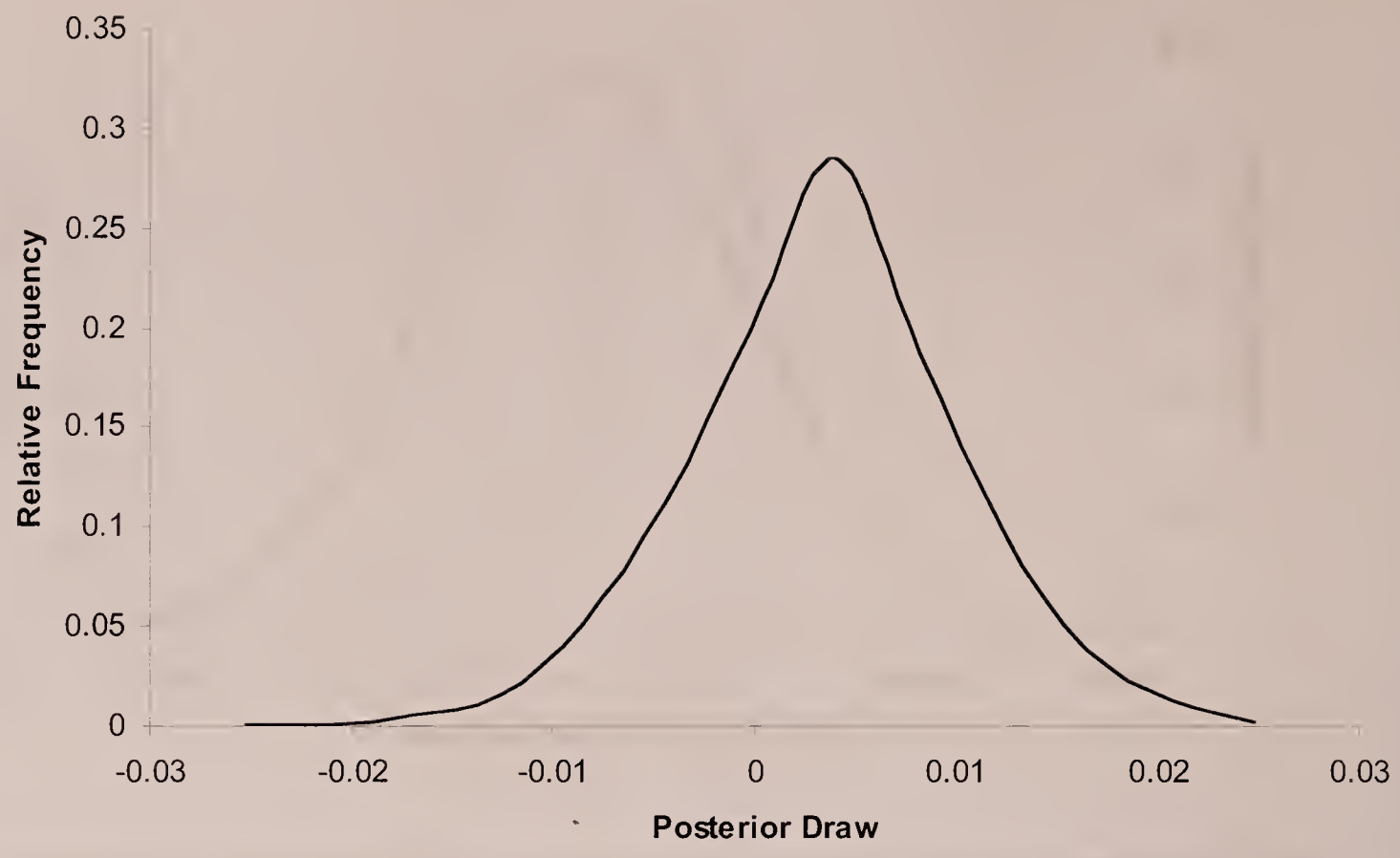

Figure 4.44. Posterior Distribution of Coefficient of Covariate Vignette Word Count for b-parameter

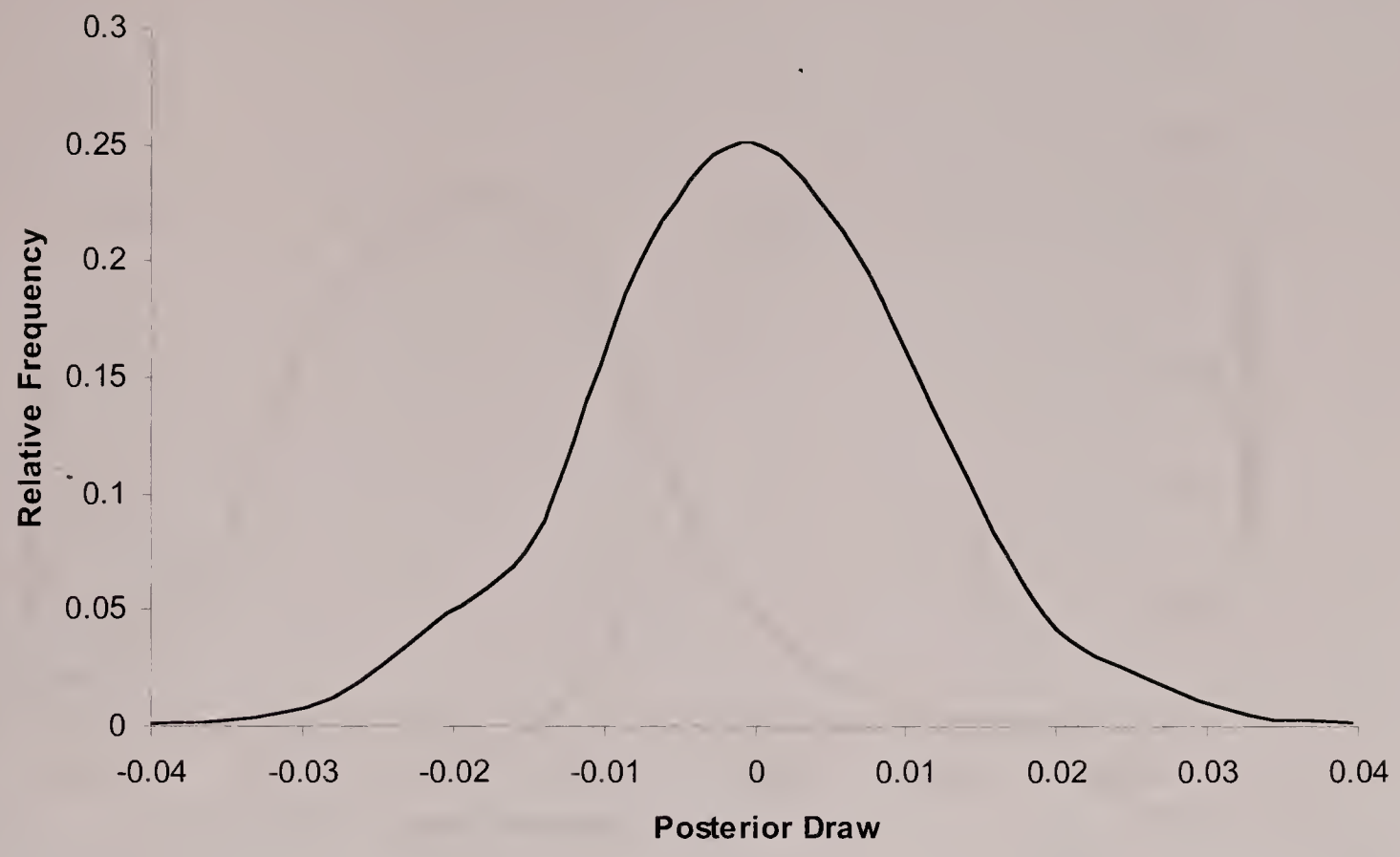

Figure 4.45. Posterior Distribution of Coefficient of Covariate Stem Word Count for b-parameter

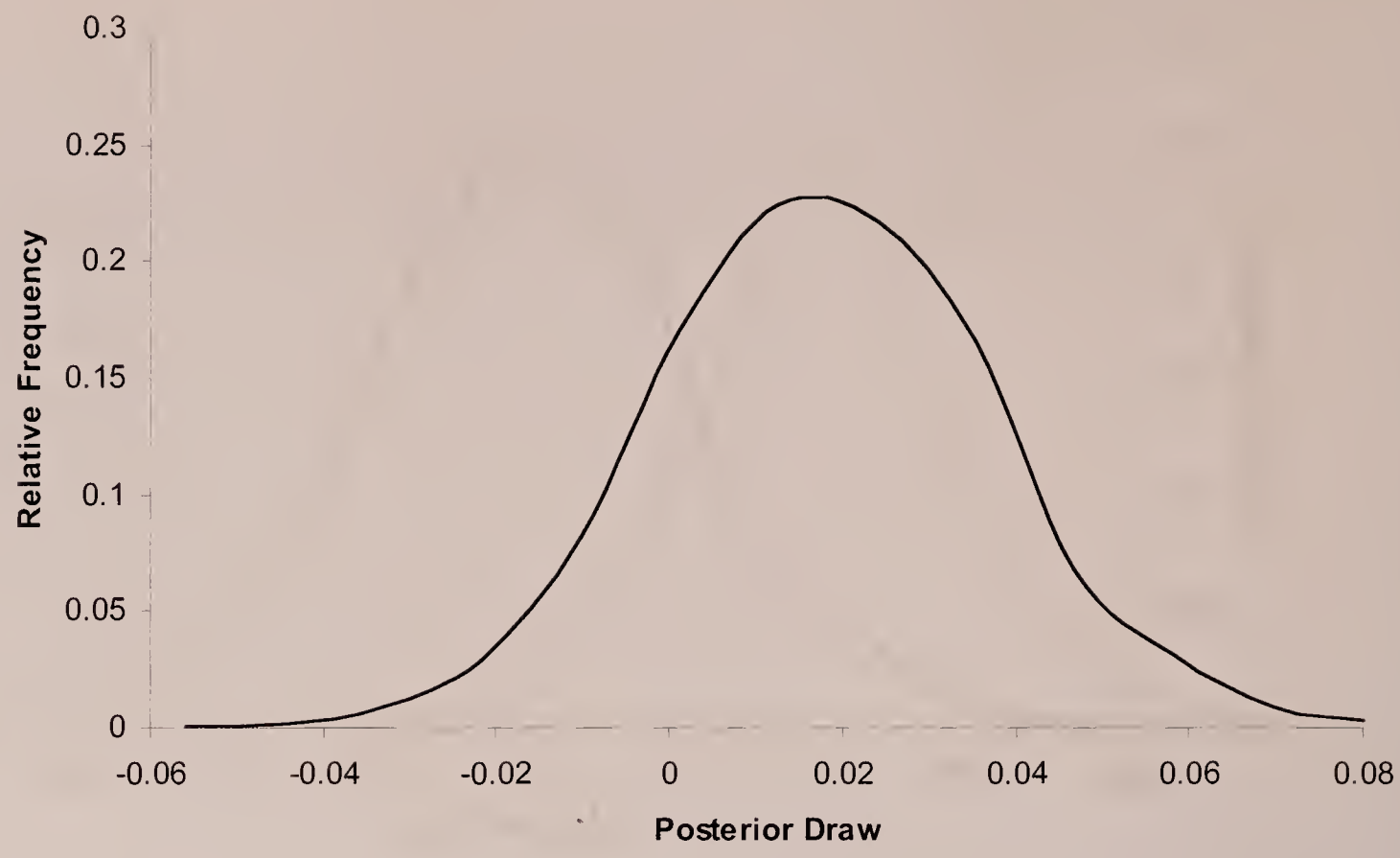

Figure 4.46. Posterior Distribution of Coefficient of Covariate Options Word Count for b-parameter

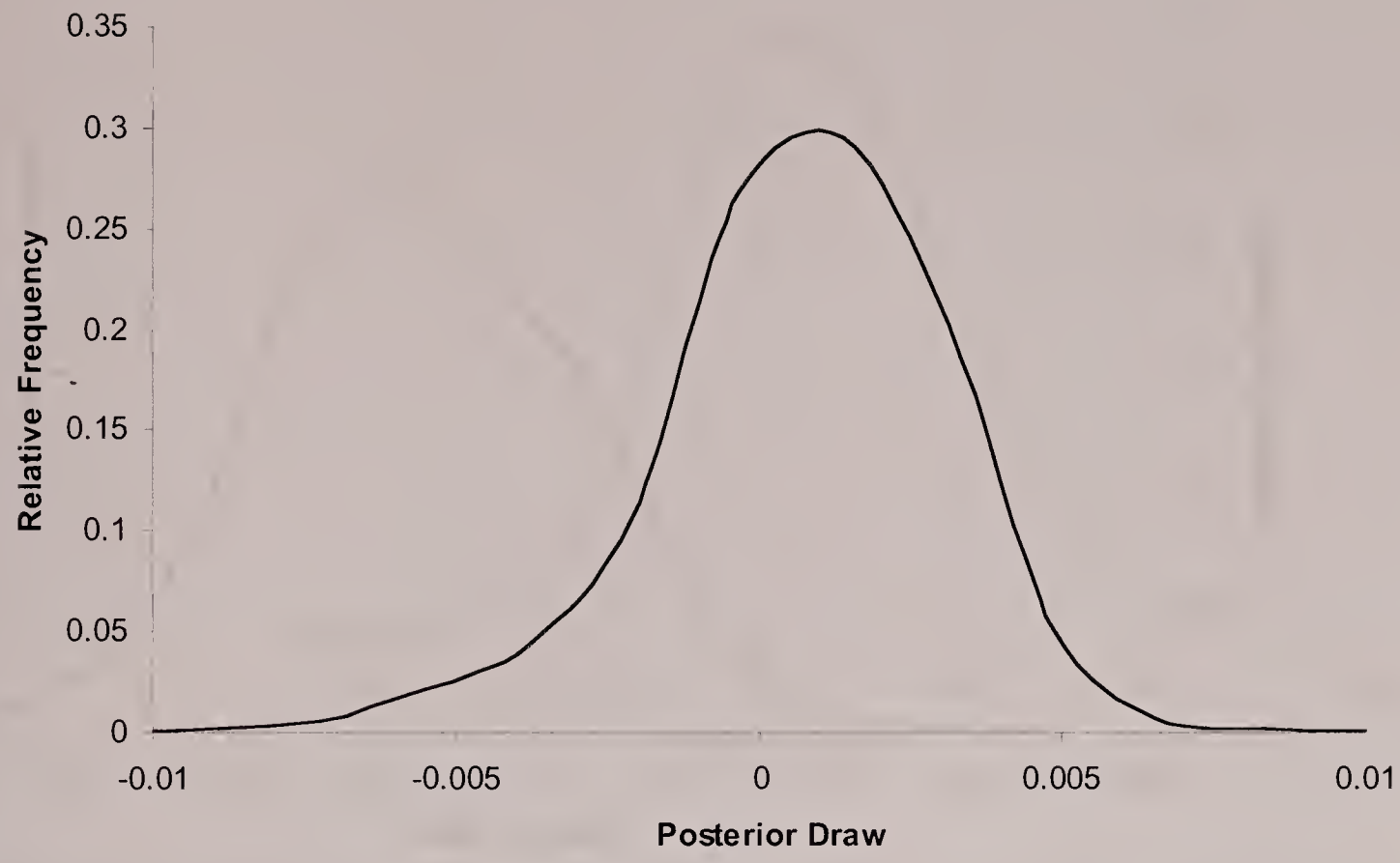

Figure 4.47. Posterior Distribution of Coefficient of Covariate Vignette Word Count for γ-parameter

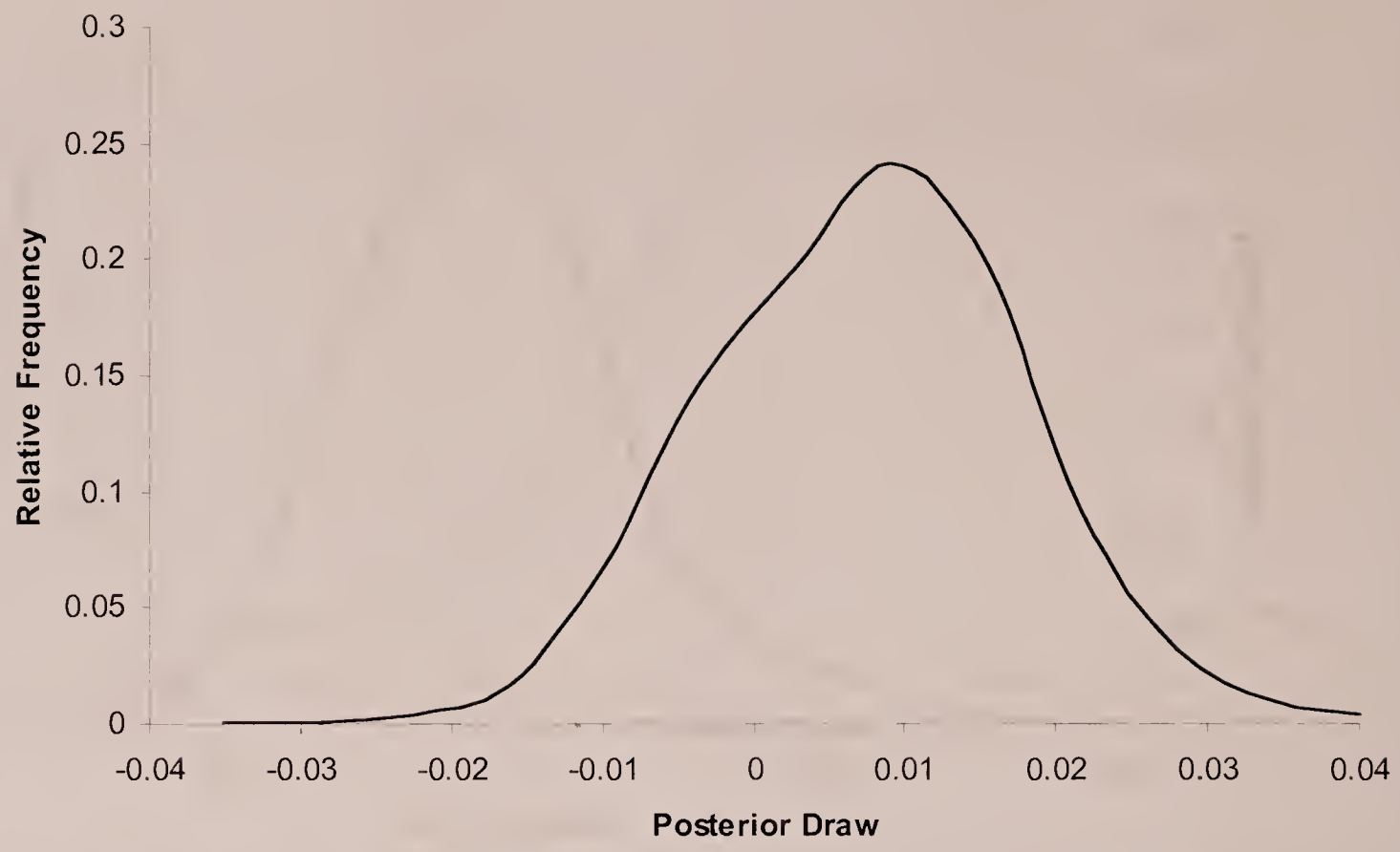

Figure 4.48. Posterior Distribution of Coefficient of Covariate Stem Word Count for γ-parameter

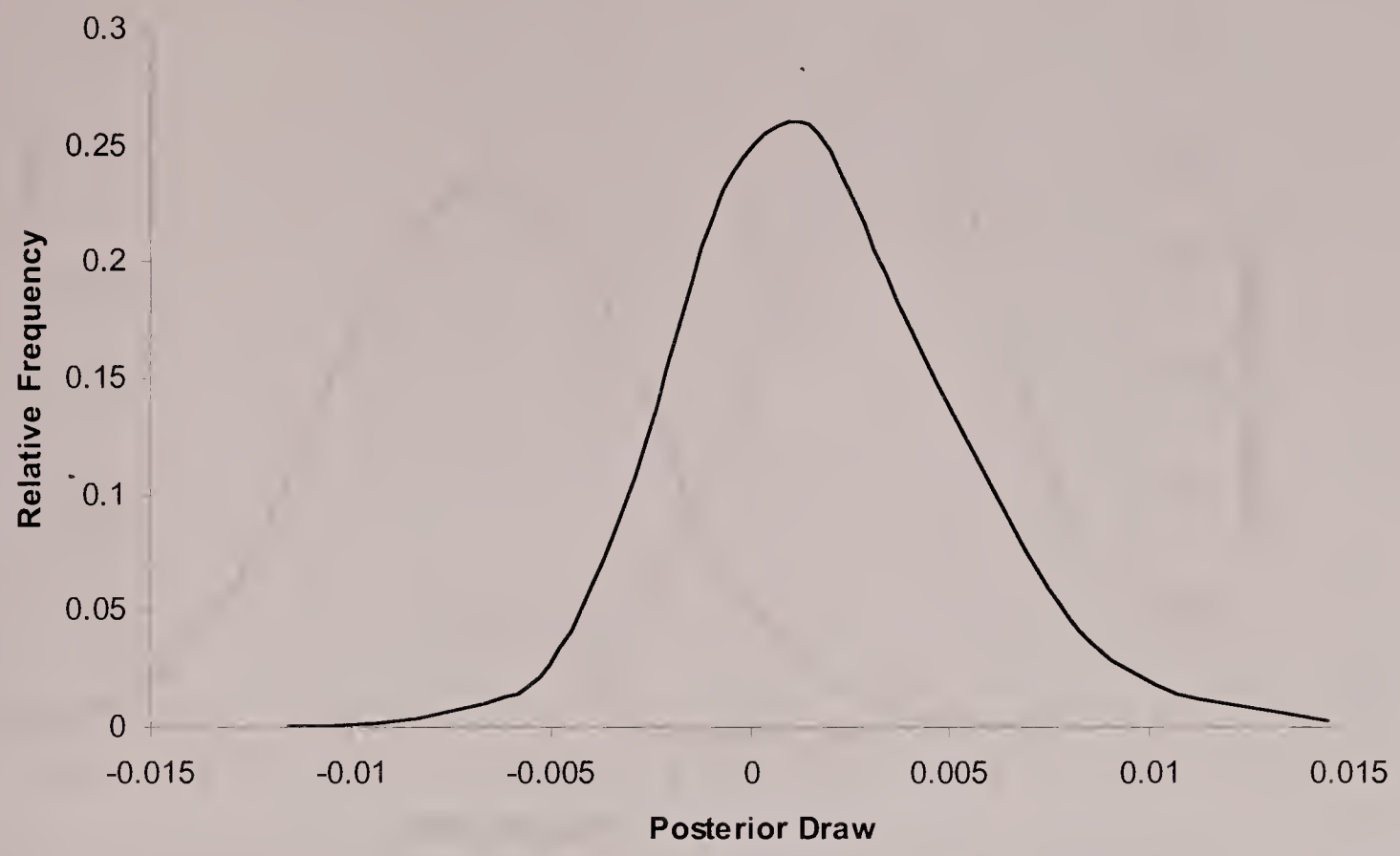

Figure 4.49. Posterior Distribution of Coefficient of Covariate Options Word Count for γ-parameter

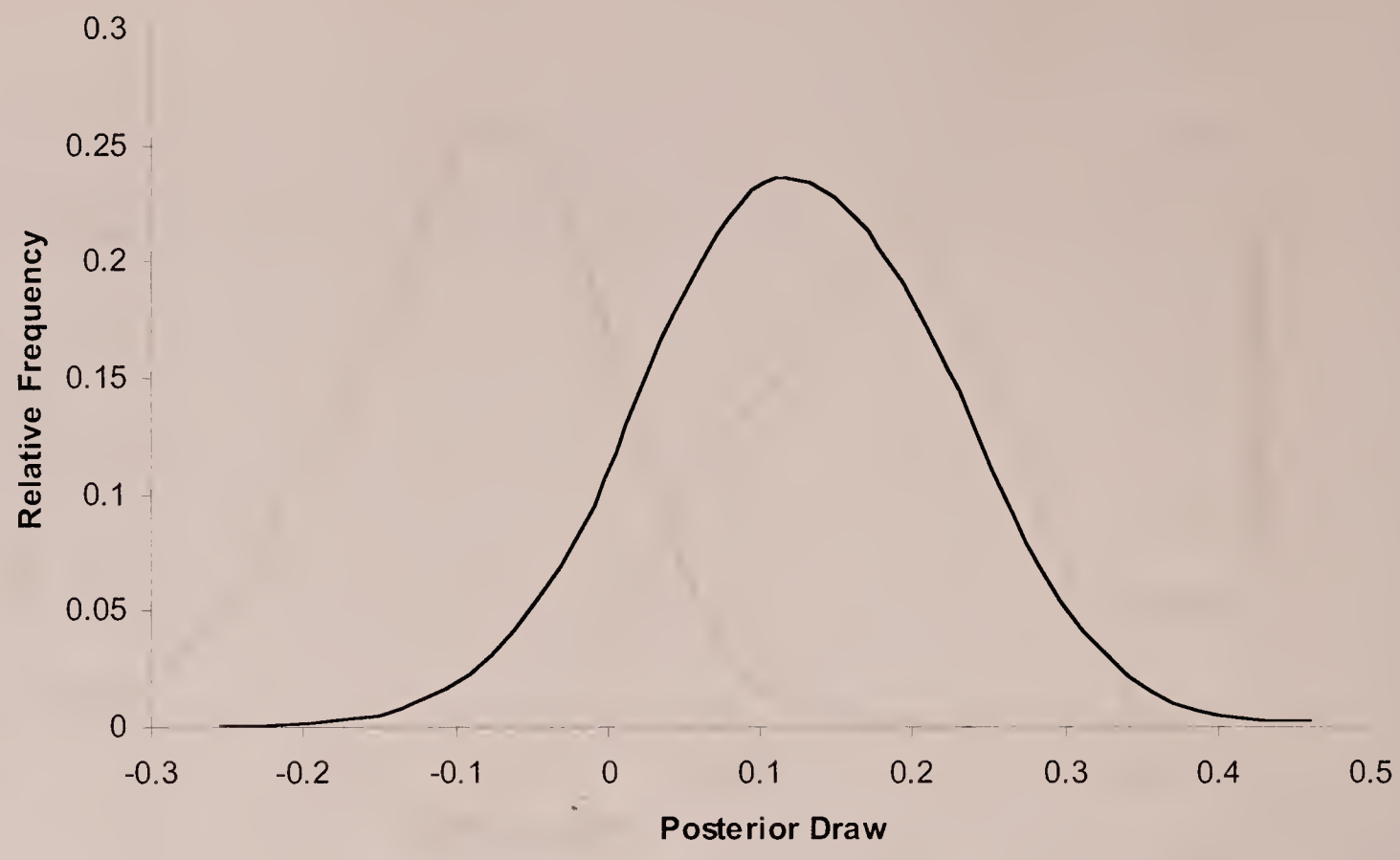

Figure 4.50. Posterior Distribution of Coefficient of Covariate Gender

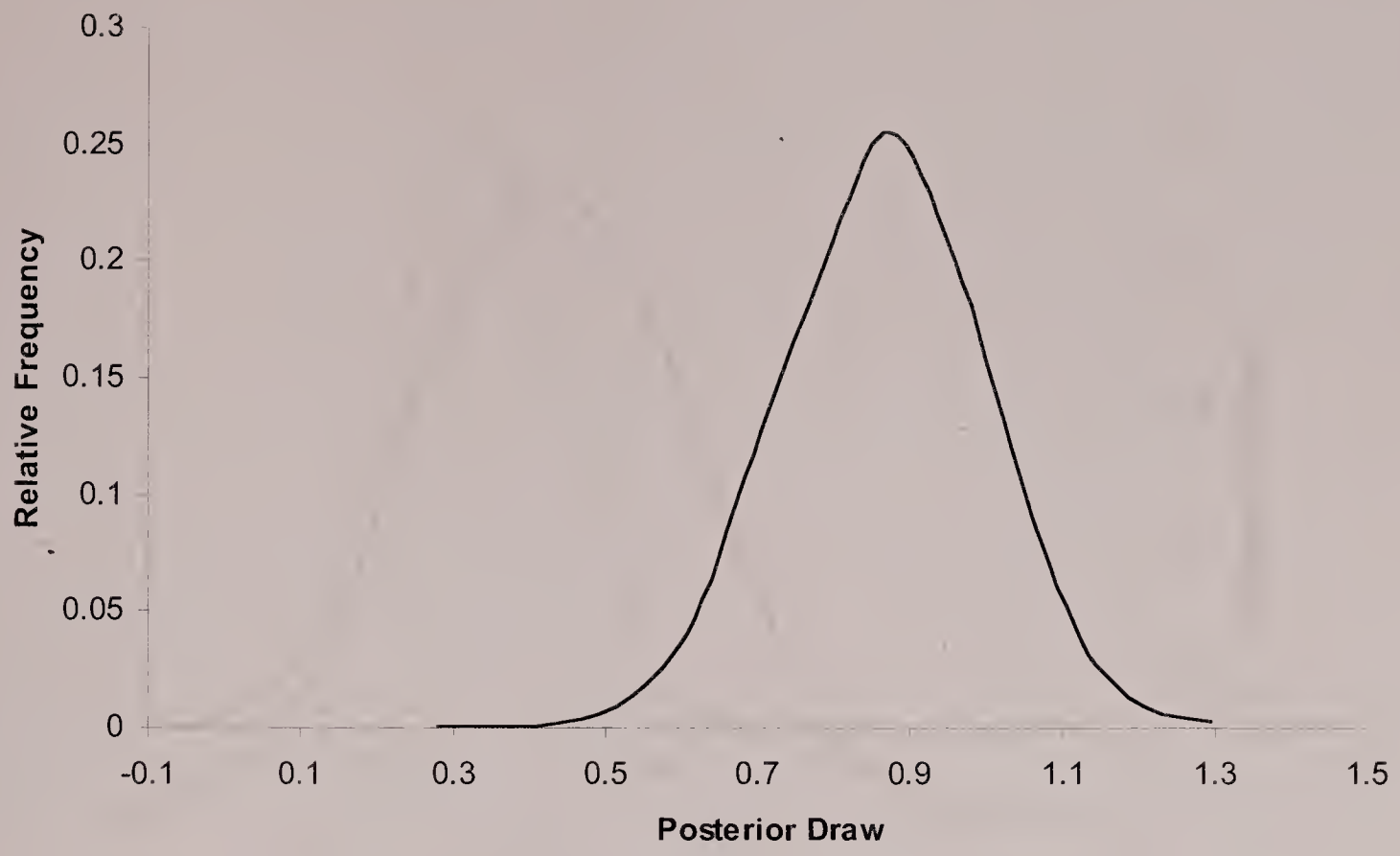

Figure 4.51. Posterior Distribution of Coefficient of Covariate LCME Status

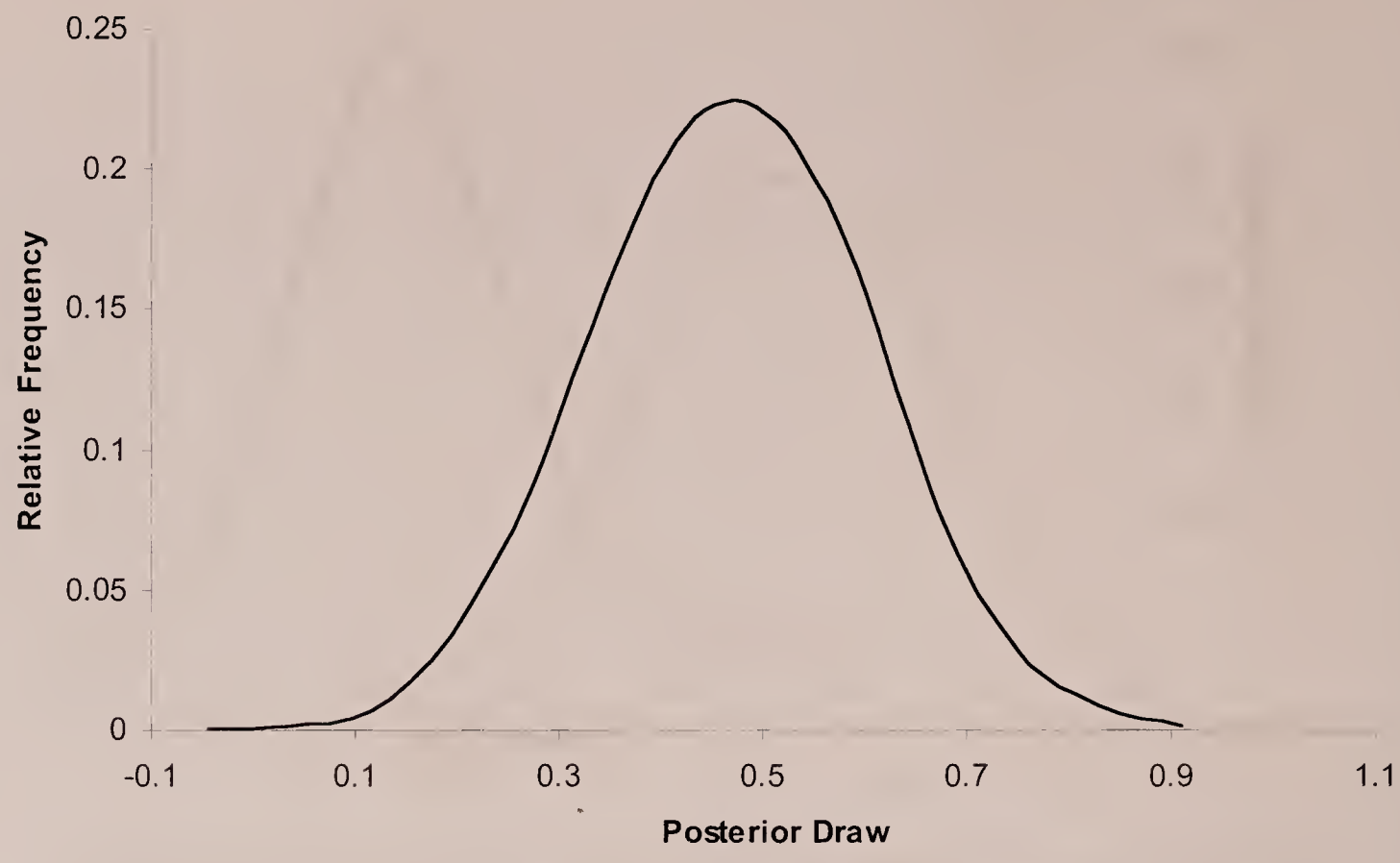

Figure 4.52. Posterior Distribution of Coefficient of Covariate Native English Speaker

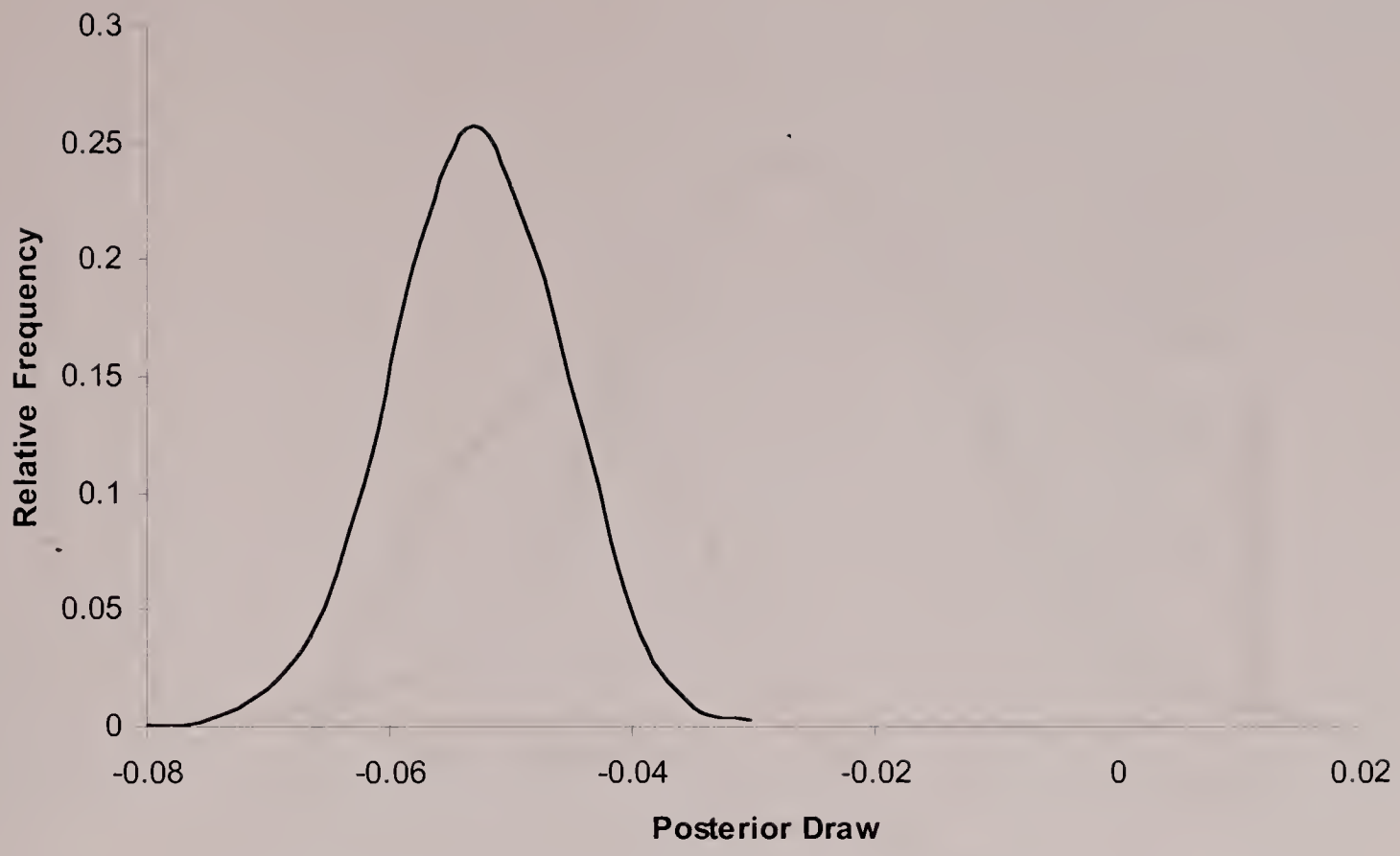

Figure 4.53. Posterior Distribution of Coefficient of Covariate Response Time

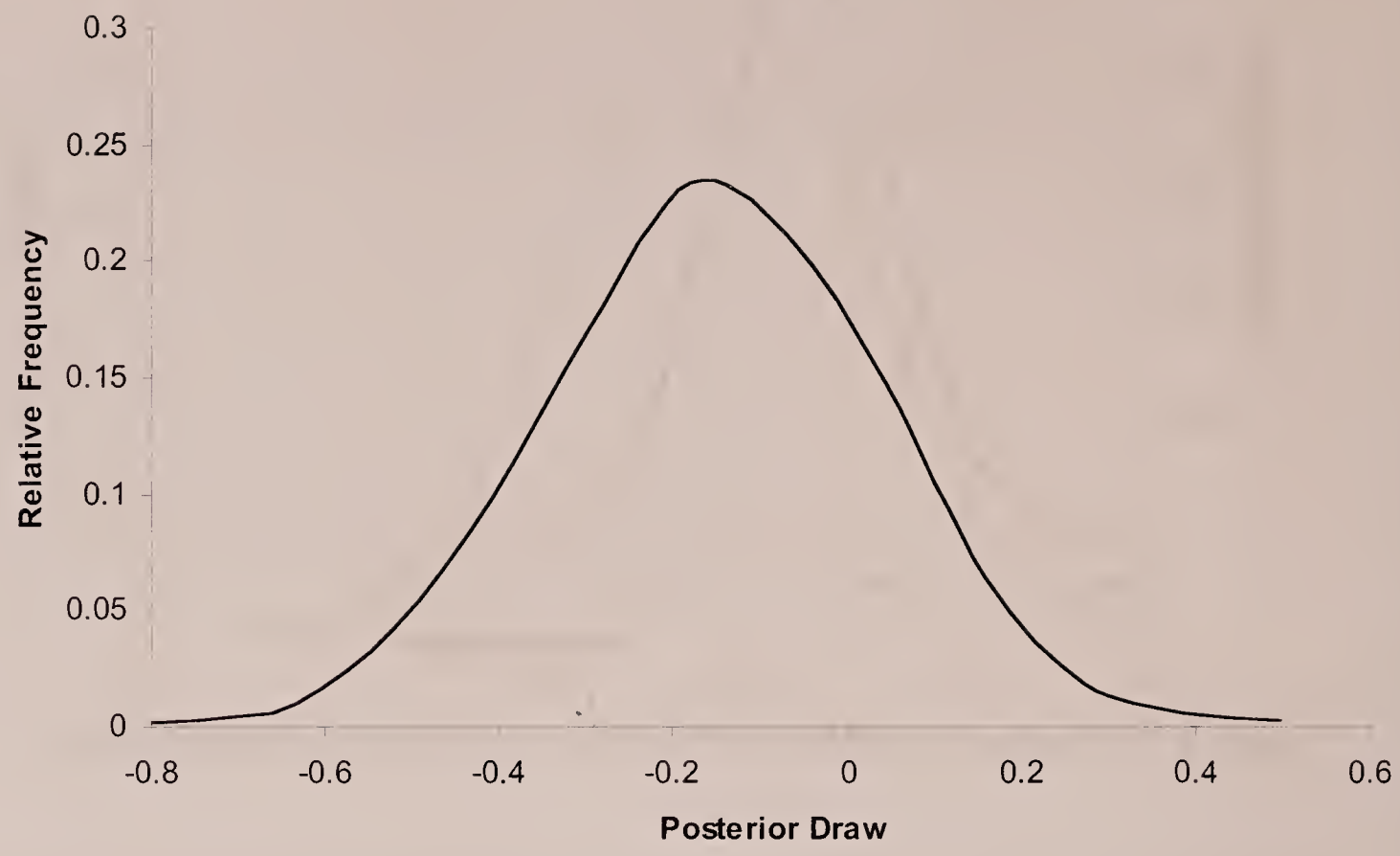

Figure 4.54. Posterior Distribution of Coefficient of Covariate Asian

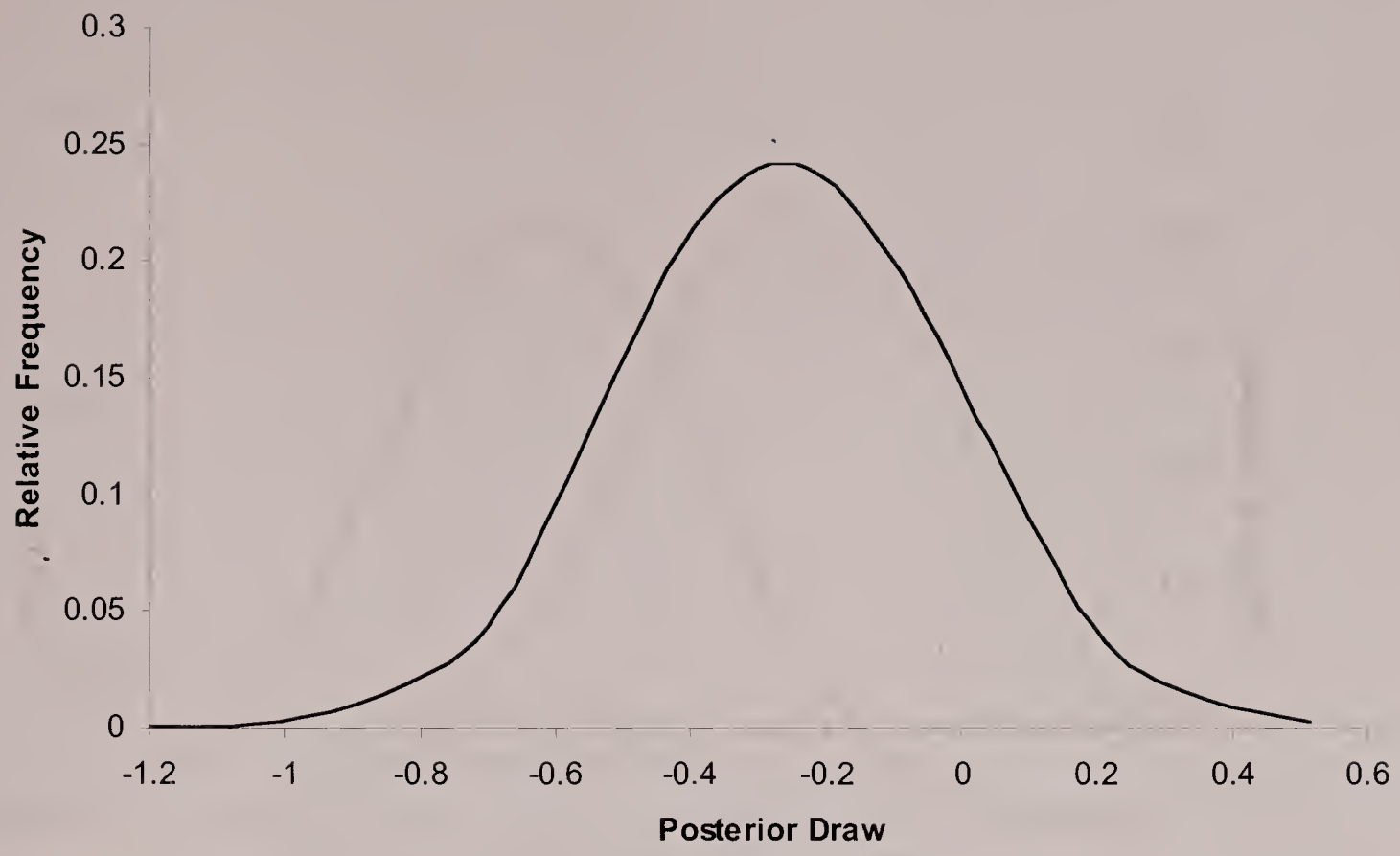

Figure 4.55. Posterior Distribution of Coefficient of Covariate Hispanic

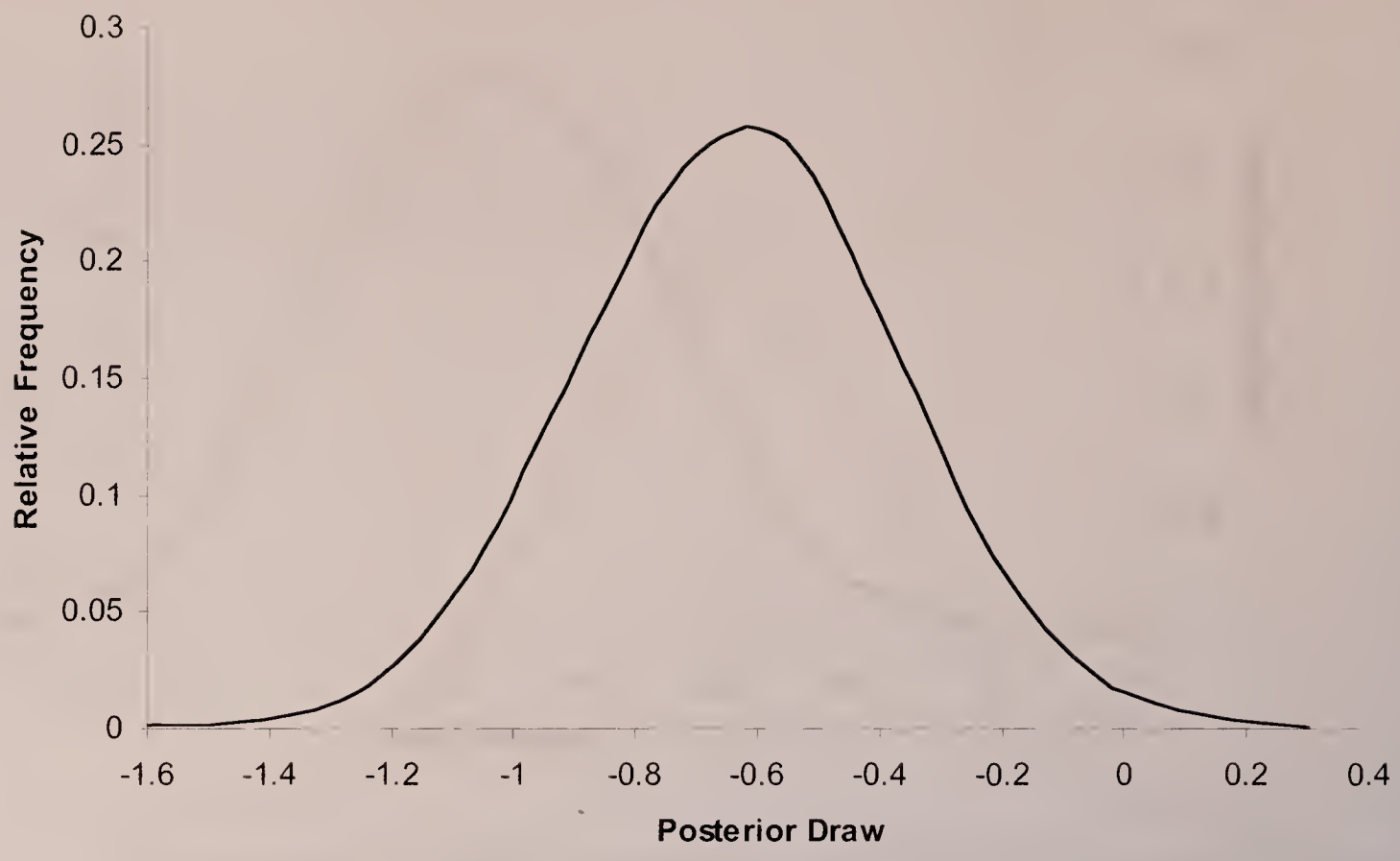

Figure 4.56. Posterior Distribution of Coefficient of Covariate Black

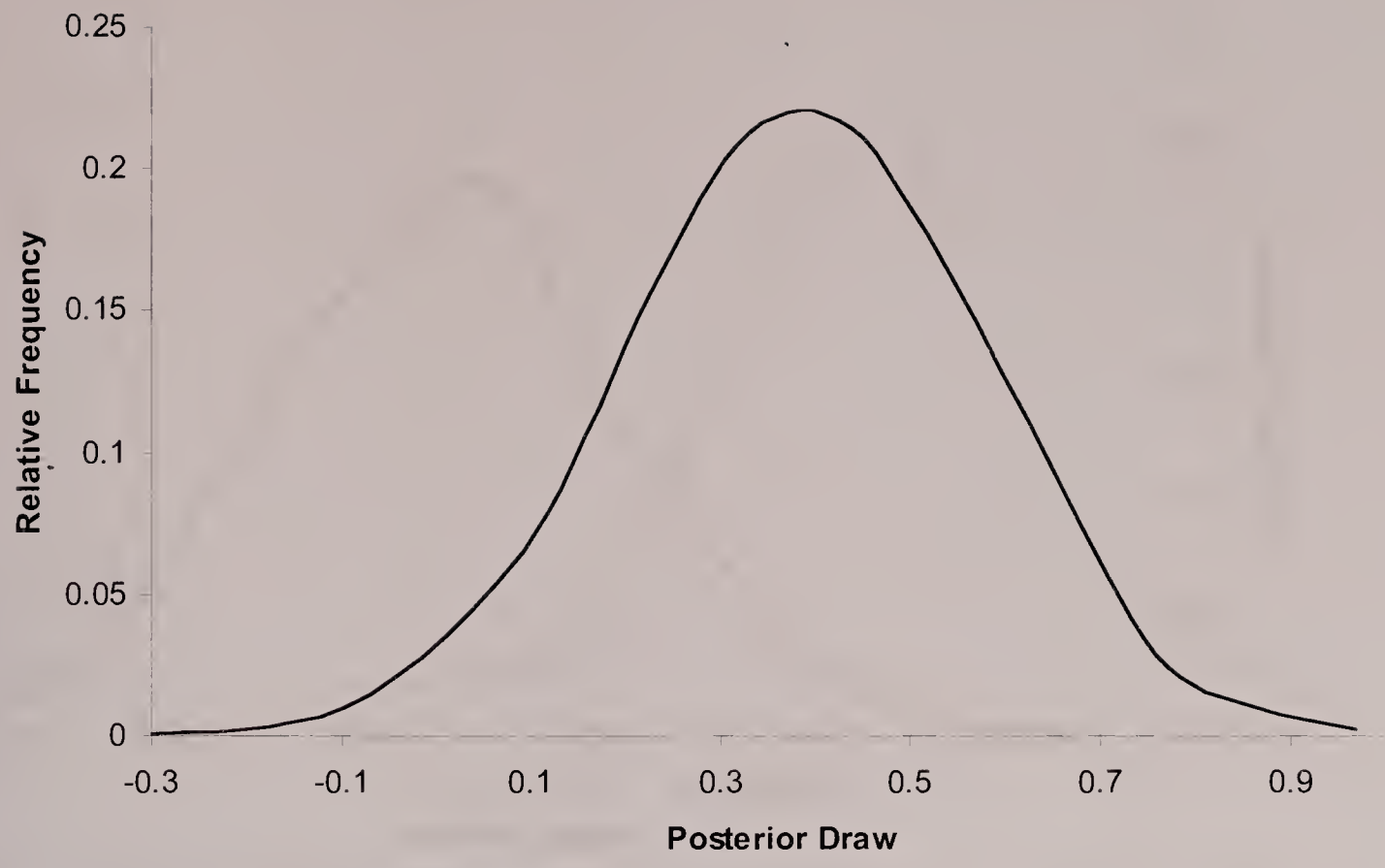

Figure 4.57. Posterior Distribution of Coefficient of Covariate White

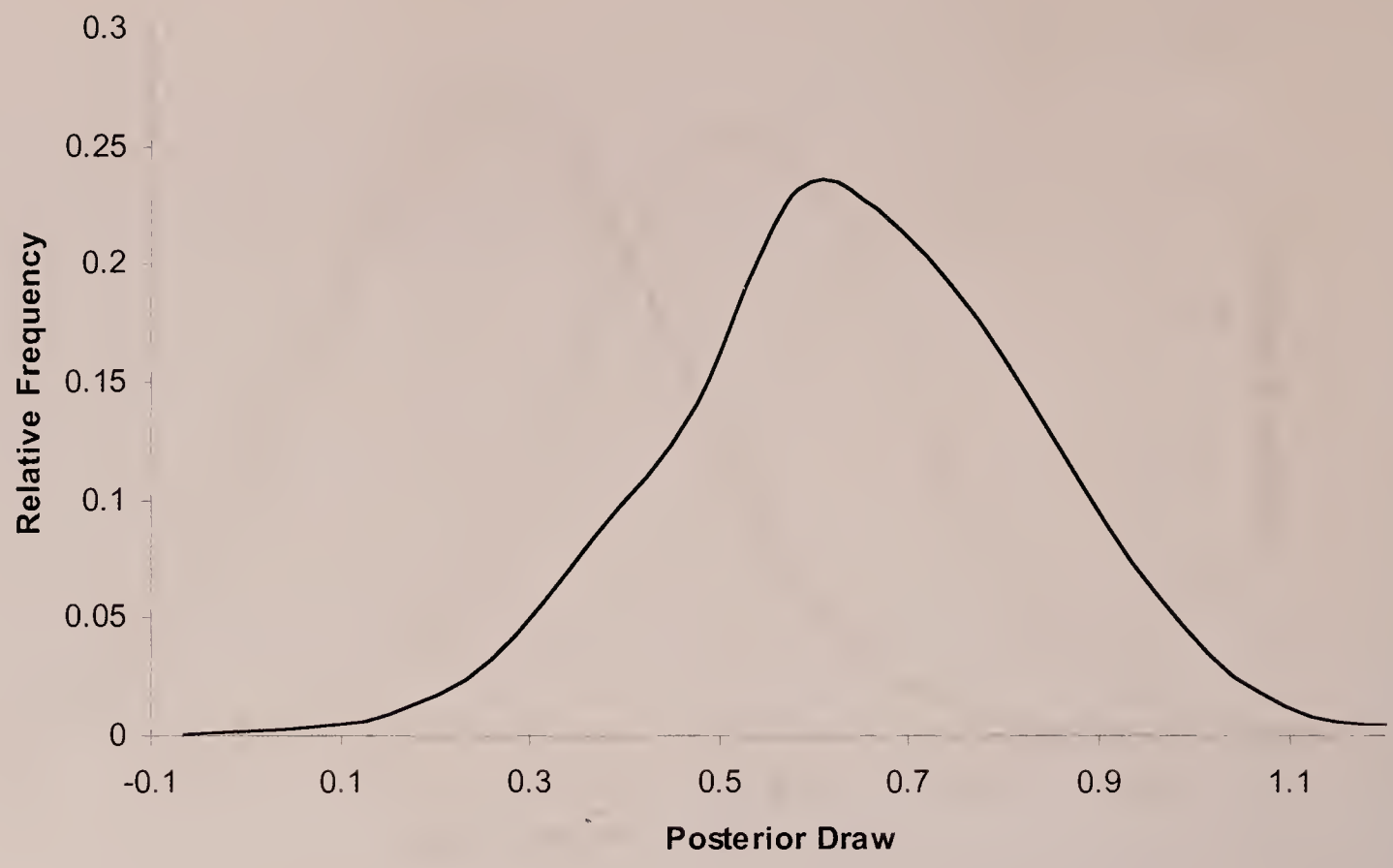

Figure 4.58. Posterior Distribution of Coefficient White minus Coefficient Hispanic

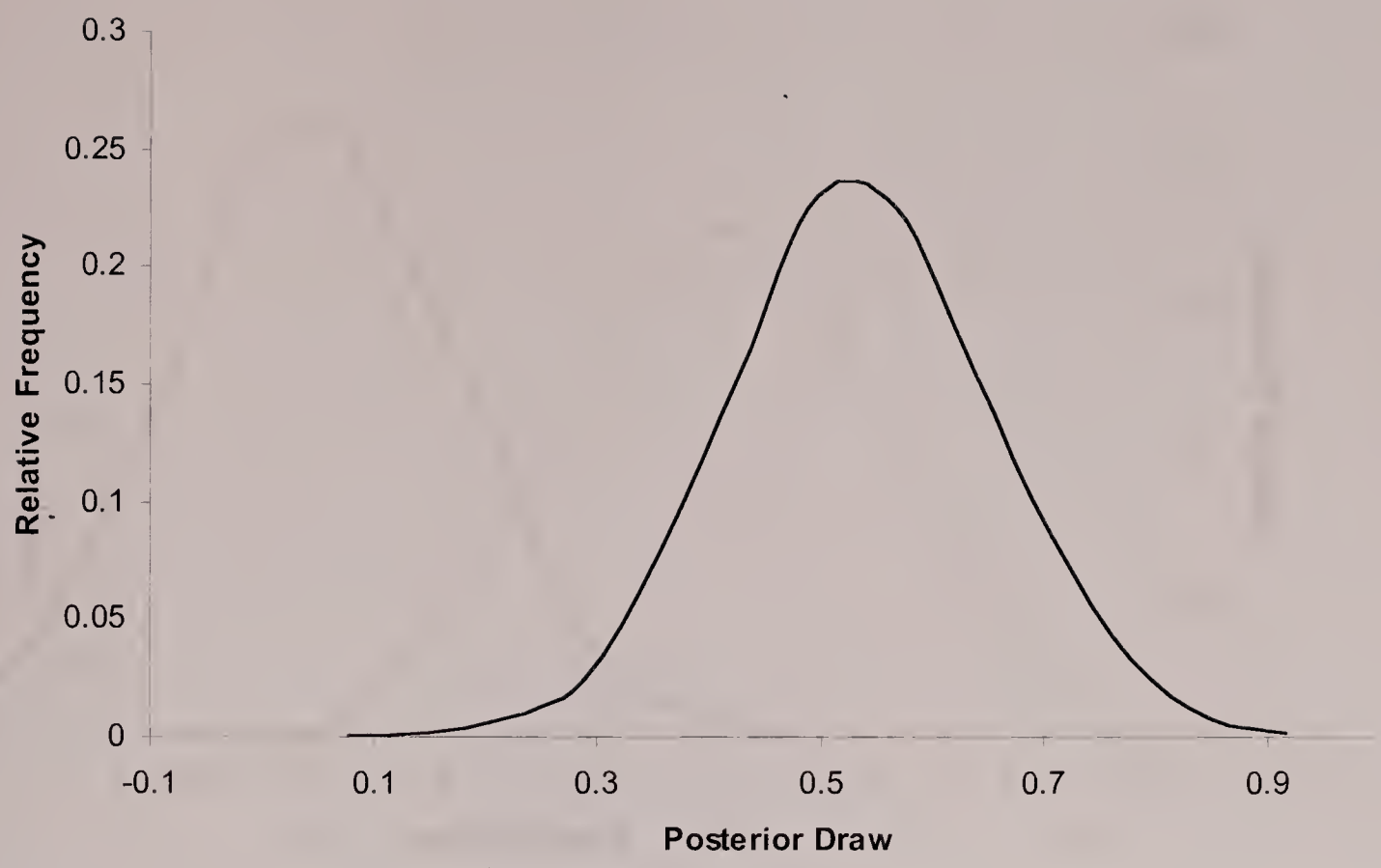

Figure 4.59. Posterior Distribution of Coefficient White minus Coefficient Asian

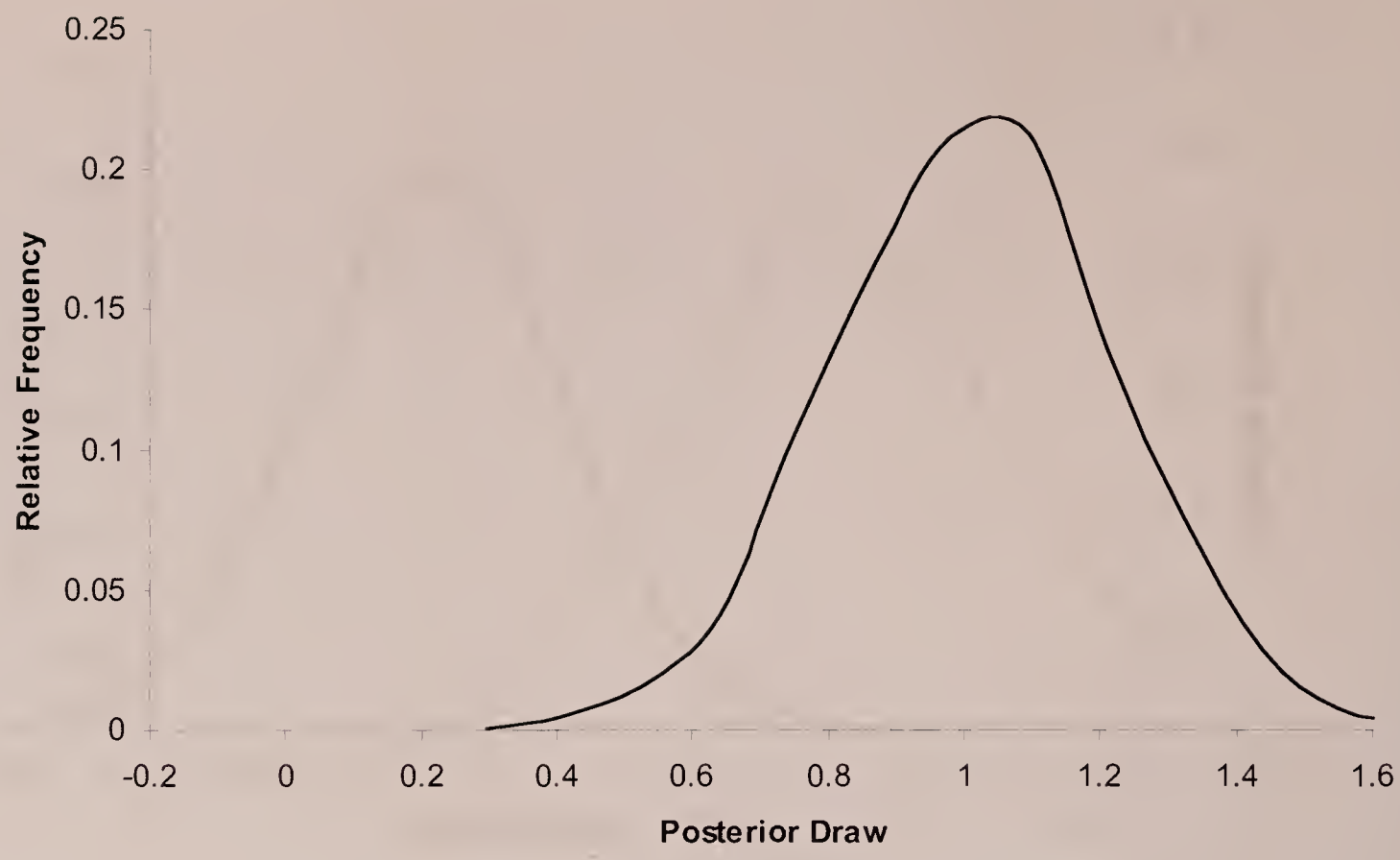

Figure 4.60. Posterior Distribution of Coefficient White minus Coefficient Black

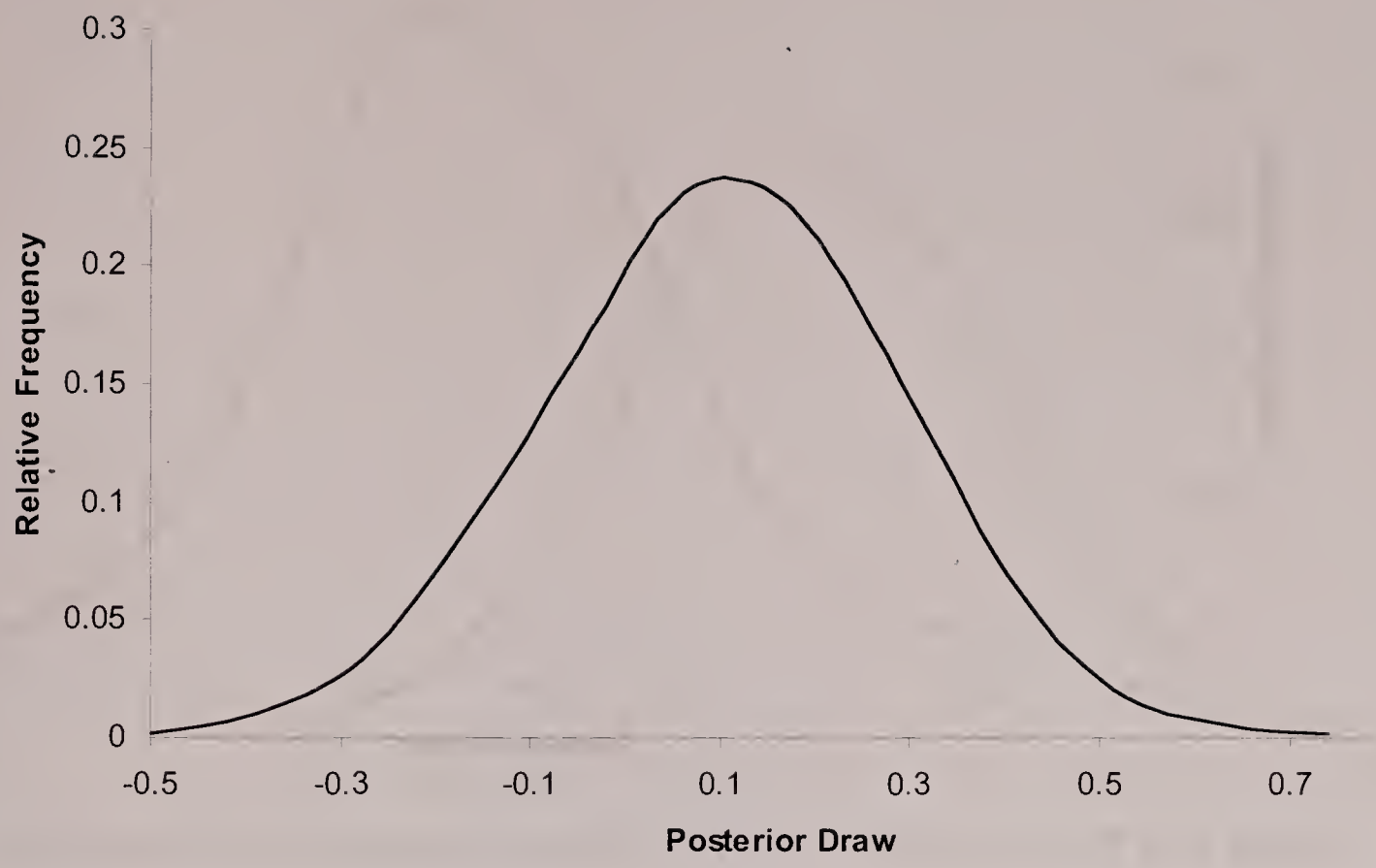

Figure 4.61. Posterior Distribution of Coefficient Asian minus Coefficient Hispanic

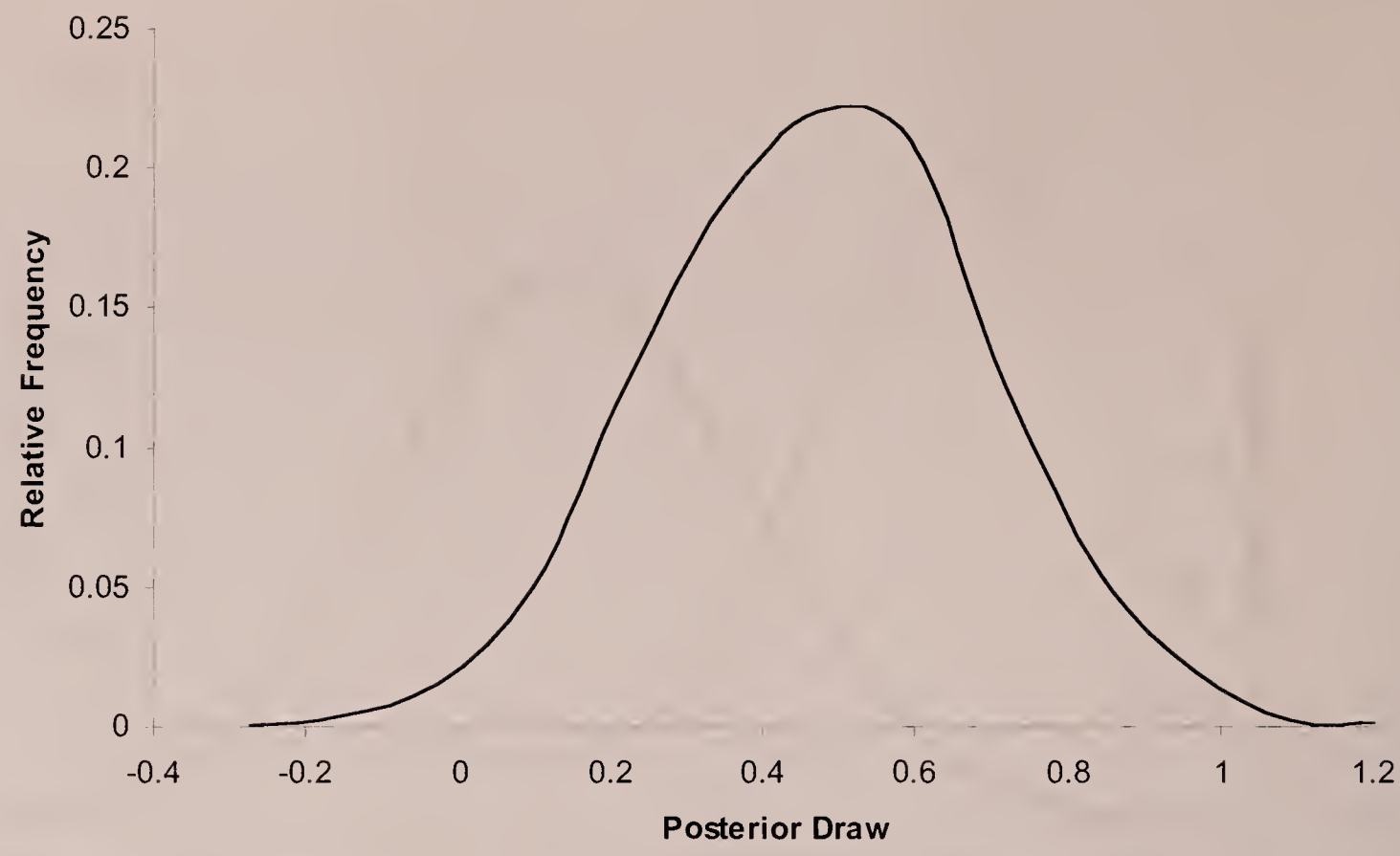

Figure 4.62. Posterior Distribution of Coefficient Asian minus Coefficient Black

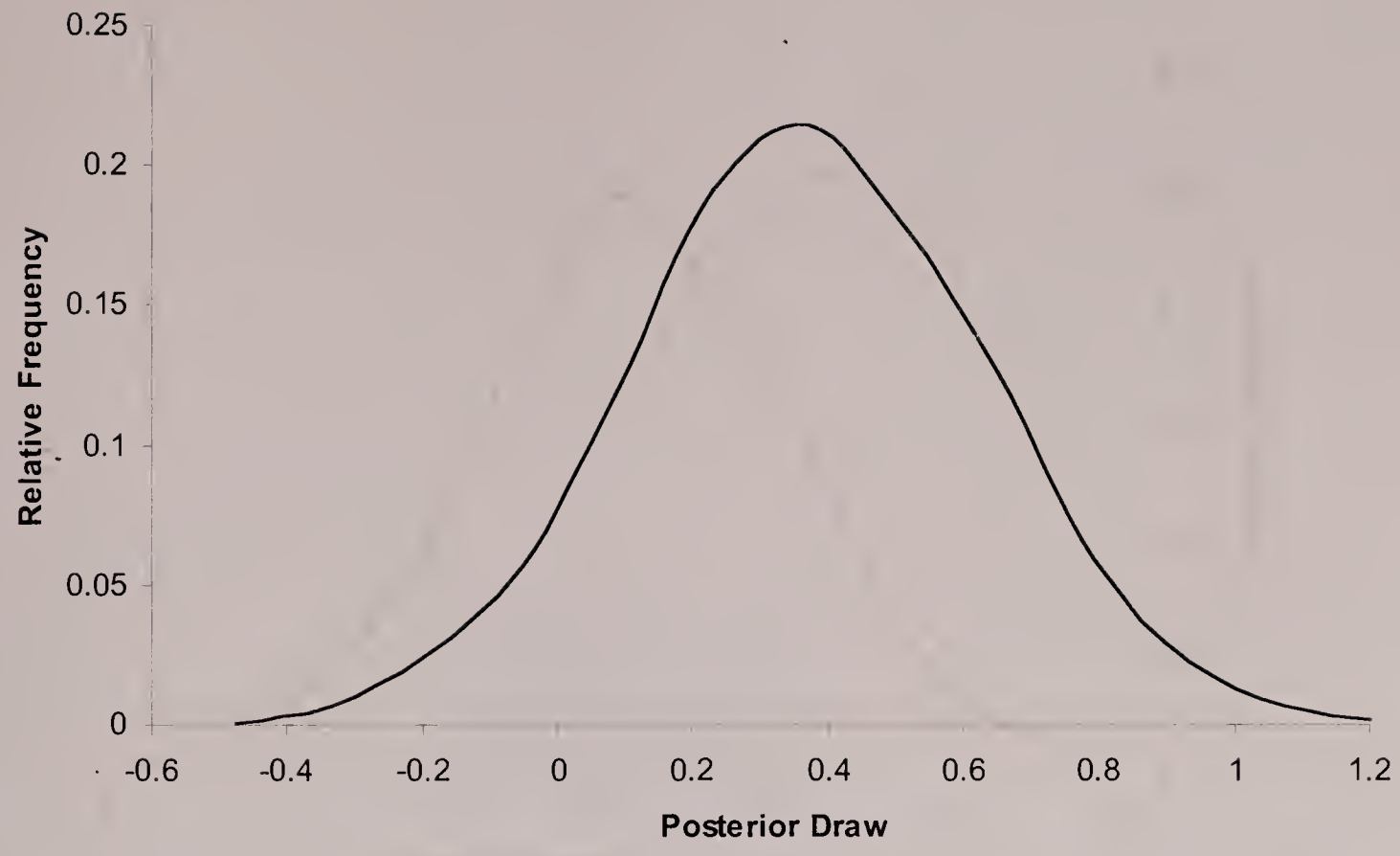

Figure 4.63. Posterior Distribution of Coefficient Hispanic minus Coefficient Black

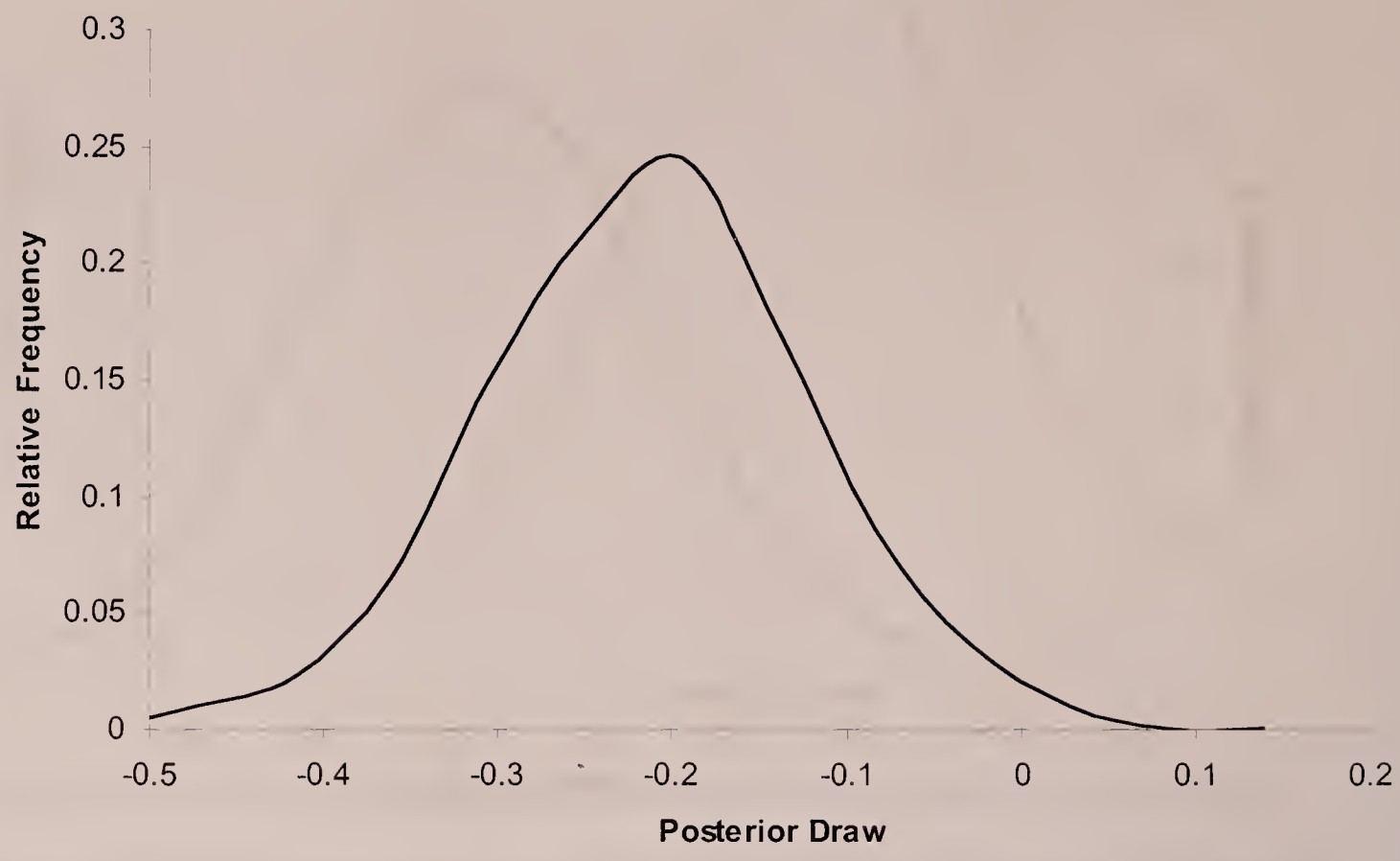

Figure 4.64. Posterior Distribution of Coefficient of Covariate White

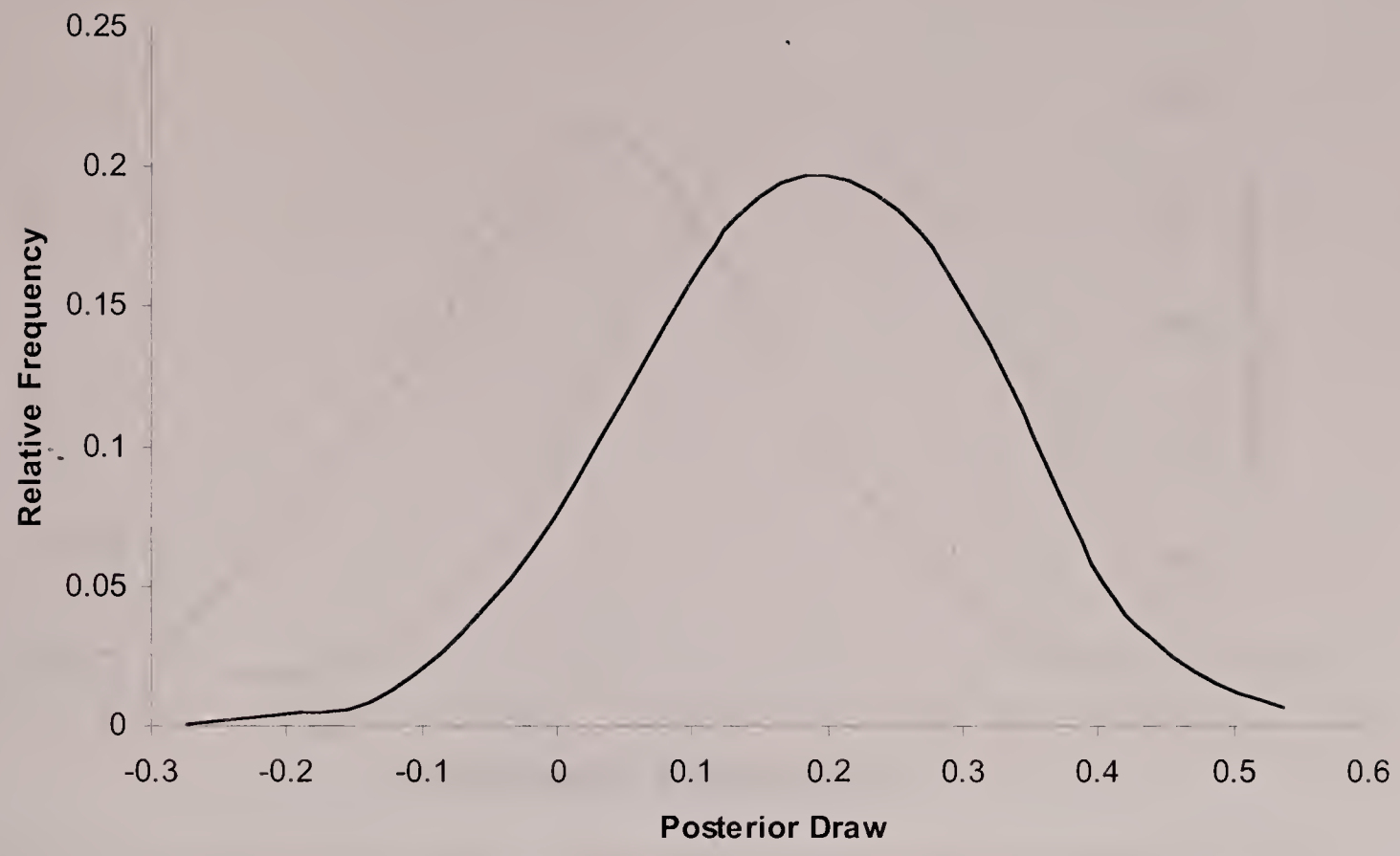

Figure 4.65. Posterior Distribution of Coefficient of Covariate Hispanic

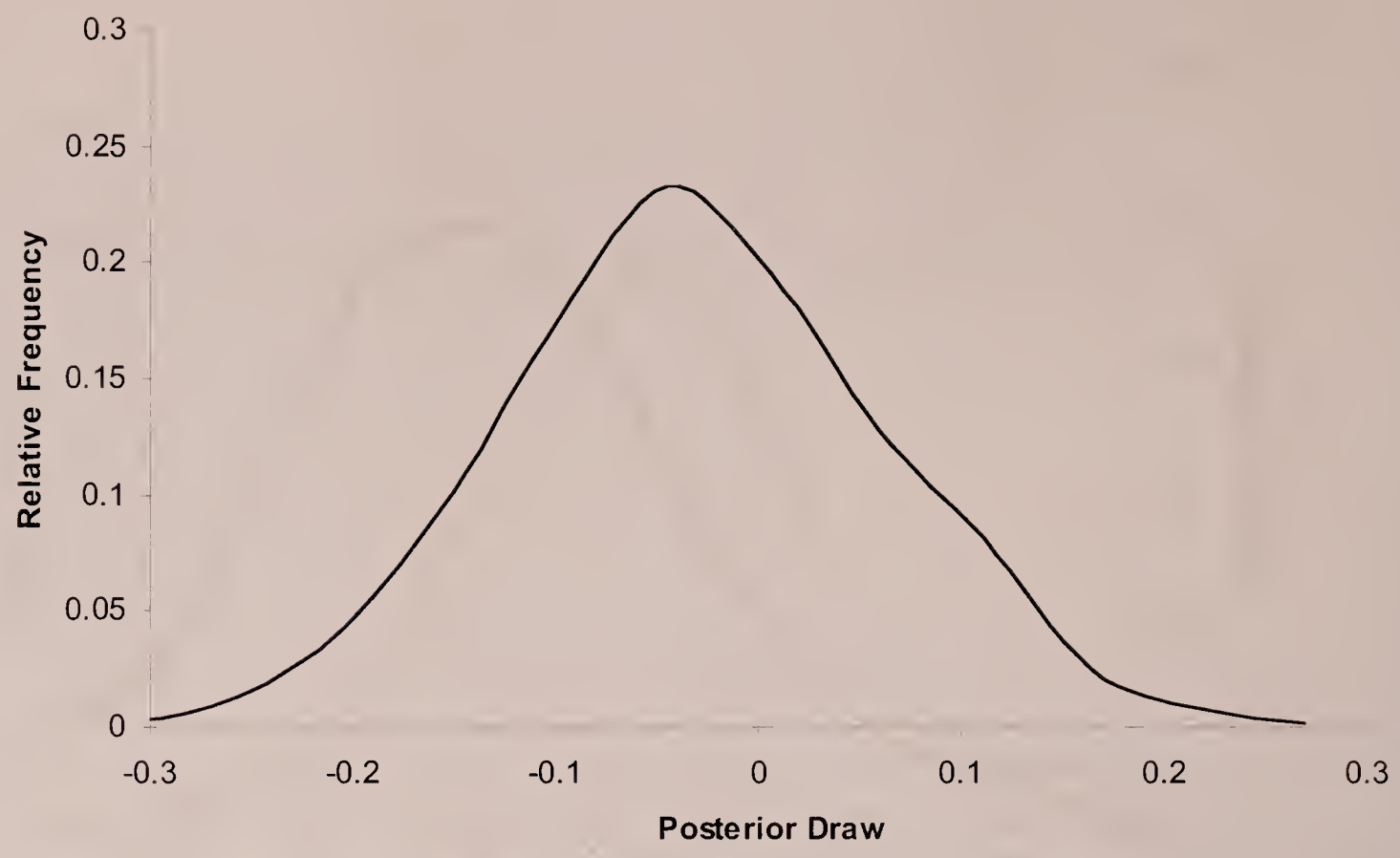

Figure 4.66. Posterior Distribution of Coefficient of Covariate Married

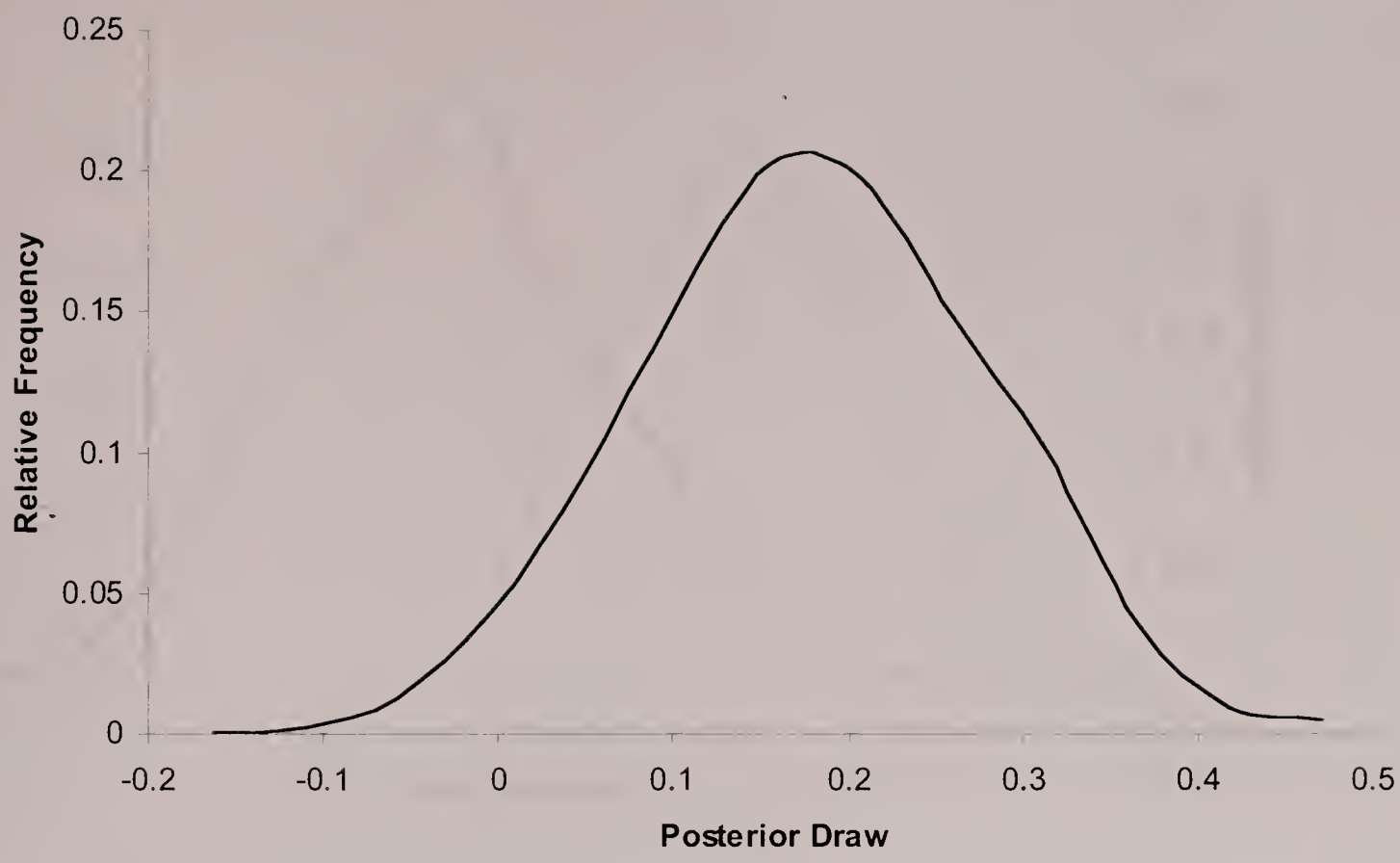

Figure 4.67. Posterior Distribution of Coefficient of Covariate Working

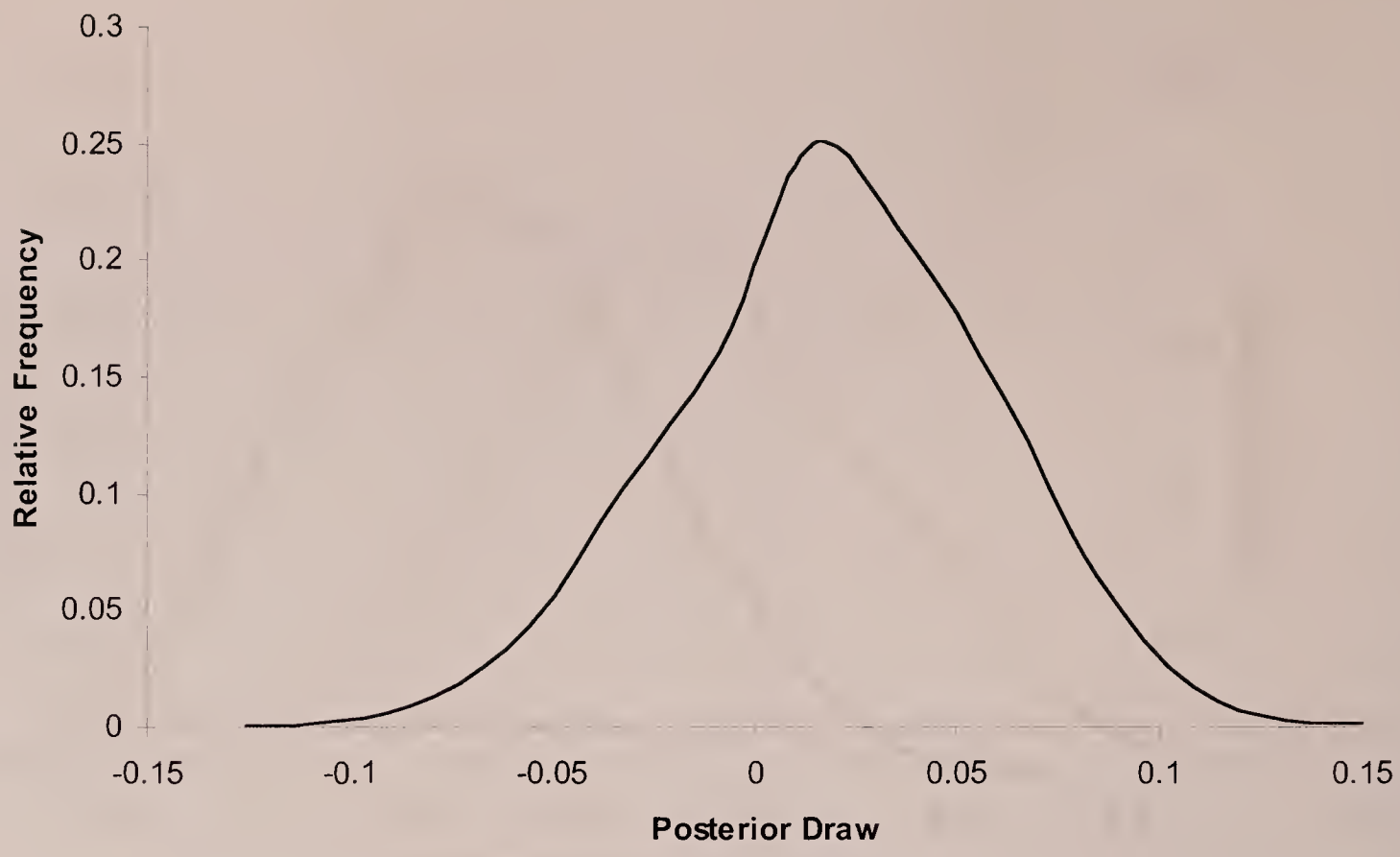

Figure 4.68. Posterior Distribution of Coefficient of Covariate Income

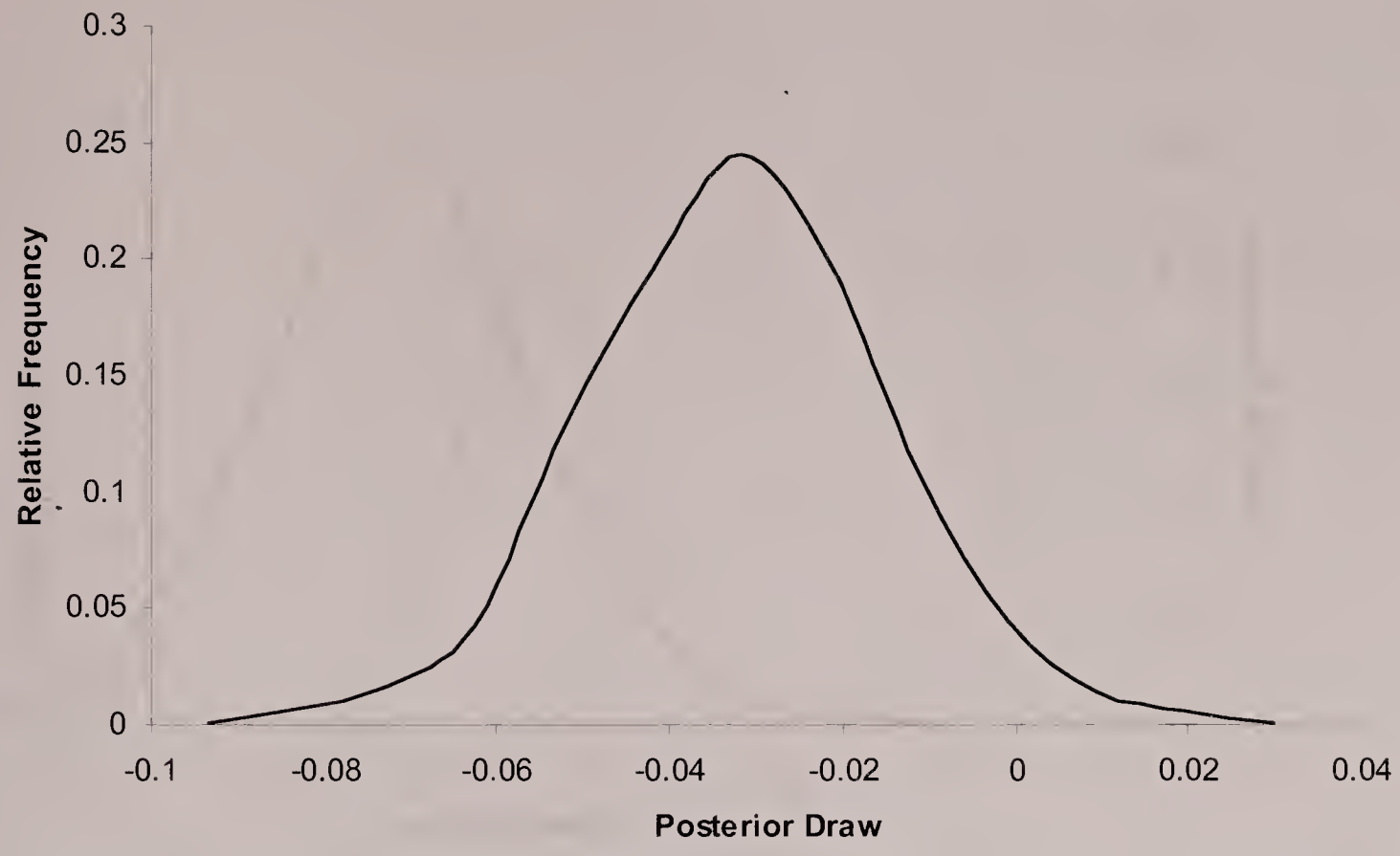

Figure 4.69. Posterior Distribution of Coefficient of Covariate Months Since Diagnosis

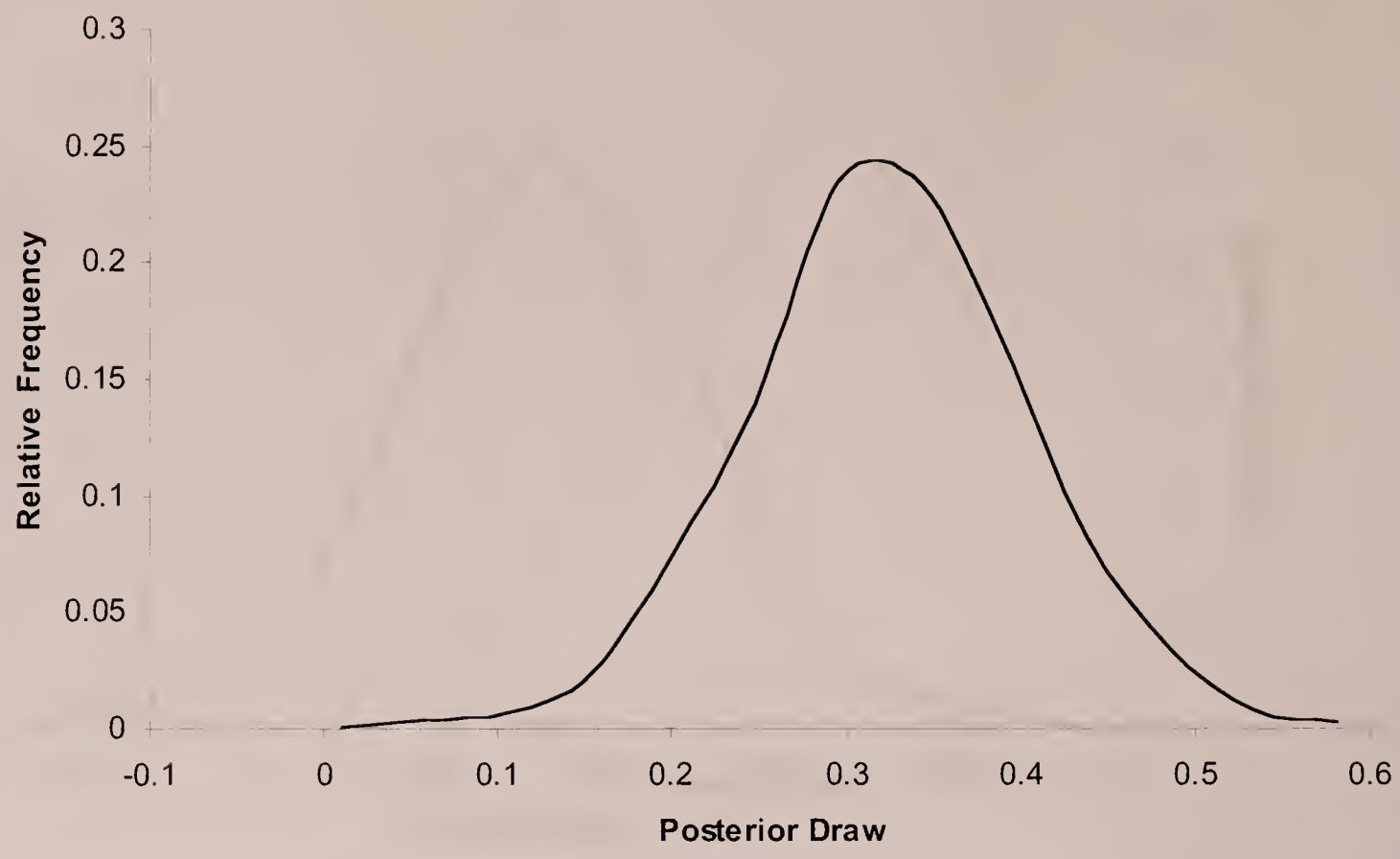

Figure 4.70. Posterior Distribution of Coefficient of Covariate Using Tamoxifen

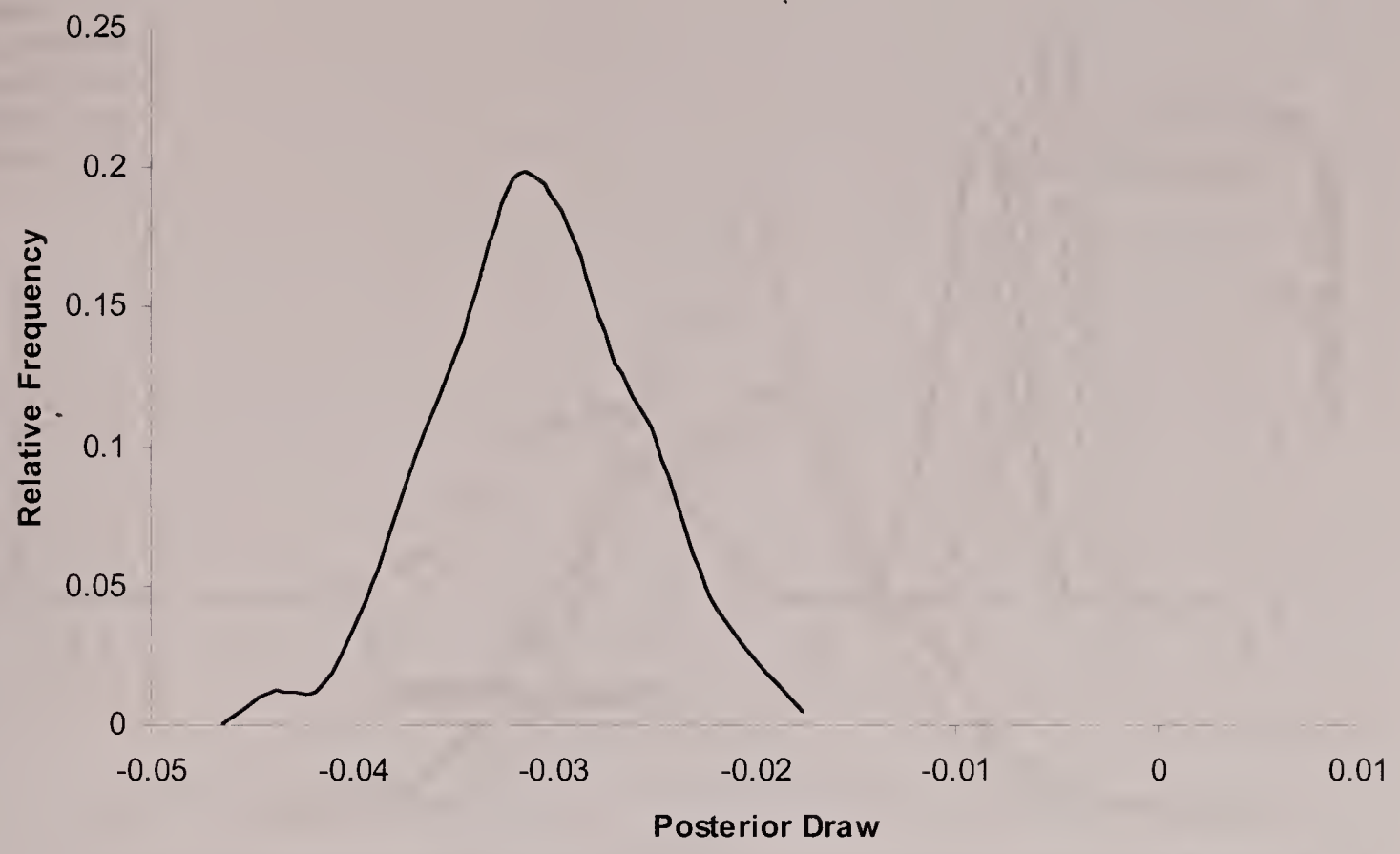

Figure 4.71. Posterior Distribution of Coefficient of Covariate Age

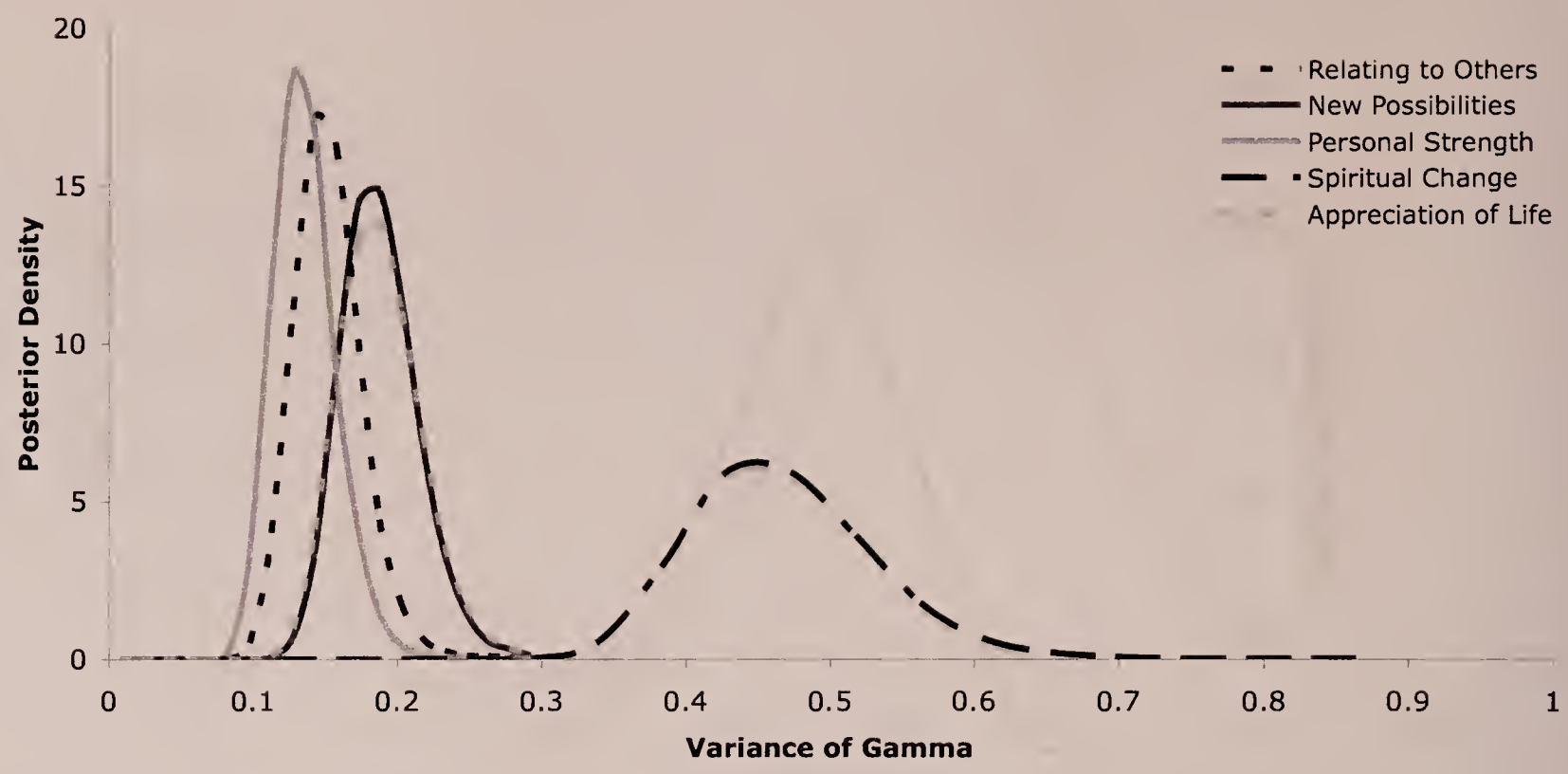

Figure 4.72. Testlet effects: Posterior Distribution of Variance of Gamma

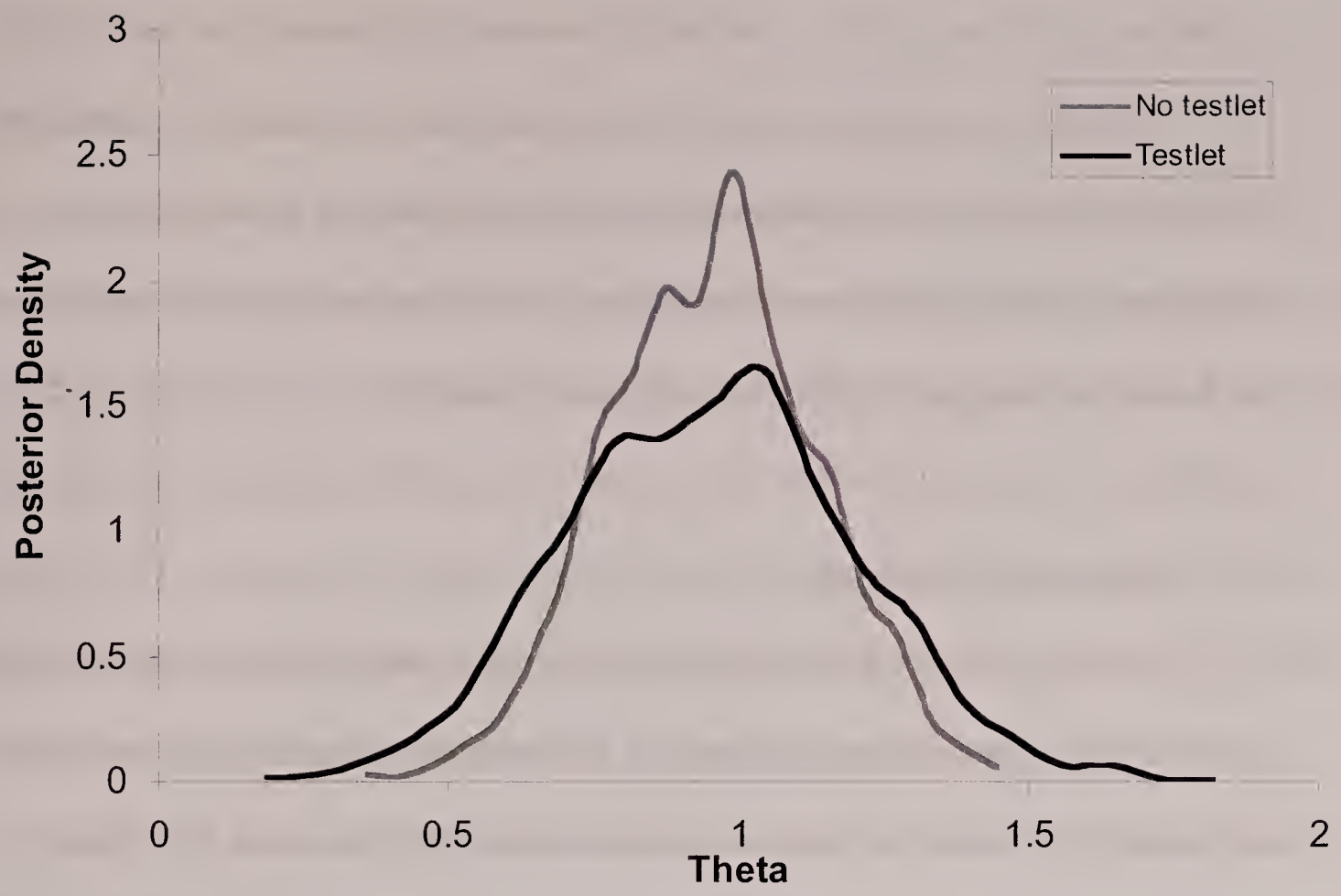

Figure 4.73. Posterior Distribution of Theta with and without the Assumption of Local Independence

CHAPTER 5

DISCUSSION

The previous chapter reported the results of the three studies conducted in this dissertation. This chapter includes a summary of the findings from these studies and a discussion of their significance. The chapter closes by presenting some of the study limitations and suggesting directions for future research.

5.1 Summary of Findings

The two primary goals of this research were: first, to conduct a basic simulation study to investigate whether the Bayesian TRT model is functioning as expected in estimating the relationship between covariates and their respective IRT model parameters; and second, to apply the Bayesian model to two empirical data sets: the Step 3 component of the USMLE, and the PTGI. As a secondary goal, the impact of not modeling extra dependency caused by testlets was also investigated for the PTGI data.

5.1.1 Simulation Study

Before examining the success of the post-hoc regression approach in recovering the true covariates, the correlations between the generating (true) thetas and the thetas estimated *without* the covariates in the model were examined for both study conditions. The average correlation across 50 replications was 0.98 for Condition 1 and 0.97 for Condition 2, which indicated that estimation process had added a small amount of error to the estimates. As expected, this resulted in slightly lower correlations between the

covariates and the estimated thetas compared to the correlations between the covariates and the true thetas. Therefore, it was not surprising that the results also showed that the post-hoc approach consistently underestimated the covariate coefficients in both conditions.

Additionally, confidence intervals for each centered decile were computed for each of the 50 replication's covariate coefficients to evaluate the success of the post-hoc regression approach in estimating the confidence in its own estimates of beta. Using these intervals, the proportion of times the true covariate fell within the given confidence interval was examined. For Condition 1, the post-hoc regression approach tended to be overconfident in its estimates of the betas, yielding smaller intervals than it should have for the majority of the deciles. However, it should be noted that the tendency of overconfidence reversed for the upper intervals, which are generally of greatest interest. One possible explanation for this tendency toward overconfidence is the post-hoc regression analyses' failure to account for the uncertainty (error) in theta. However, although this explanation seems plausible, the opposite trend was observed for Condition 2: the post-hoc approach intervals (except for the first two deciles) contained the true parameter more times than expected—i.e., the confidence in the beta was not high enough.

For the Bayesian approach, the correlations between the generating thetas and the estimated thetas from the runs with covariates were examined. The average correlation across 50 replications was 0.97 for both Condition 1 and Condition 2. And again, the correlations between the covariates and the estimated thetas were compared to the correlations between the true theta and the covariates. As opposed to the post-hoc results,

the observed correlations were higher than the true correlations on average for both conditions. One possible explanation for this bias can be found in the priors that the Bayesian method places on the thetas. Recall from chapter 3 that the prior for theta is $\theta_i \sim N(W_i\lambda, 1)$. Therefore, estimated thetas are biased toward the theta predicted by the covariates and this may result in inflated correlations between the estimated thetas and overestimated covariates.

To examine the success of the Bayesian approach in capturing the true covariates, expected coverage probability was compared to observed coverage probability for each decile using the posterior draws of covariate coefficients. The results revealed that the Bayesian approach slightly overestimates the betas. In Condition 1, the Bayesian approach resulted in more accurate recovery of the expected coverage probability for the majority of the interval deciles. However, in Condition 2, the post-hoc and Bayesian approach performed very similarly in terms of recovering the coverage probability expectations. As expected, the Bayesian approach resulted in slightly wider intervals than the post-hoc approach did for all deciles for both conditions. This result is likely due to the fact that the Bayesian model takes the uncertainty of the dependent variable (theta) into account when estimating the interval.

Lastly, the results of the RMSE and bias for the two approaches varied across covariates. On average, the Bayesian method had slightly less bias and error; however, the differences between the two approaches were extremely small.

5.1.2 Empirical Studies

The first empirical study employed the Bayesian approach to model the IRT model parameters, testlet parameters, and coefficients of the model parameter covariates for the USMLE Step 3. The results suggested that the item and testlet parameters were not related to the three covariates considered in the study: vignette word count, stem word count, and options word count. However, for examinees, result indicated that those from the LCME group generally have higher proficiency than those not from that group, conditional on the same values for the rest of the parameters. Native language also appears to be related to proficiency: conditional on the same values for the rest of the parameters, examinees who are native English speakers have a higher average proficiency than nonnative English speakers. Response time, *White* and *Black* group memberships were also significantly related to proficiency. The conventional post-hoc analyses were completely in line with the Bayesian findings, designating the same variables as significant as the Bayesian approach.

The second empirical study showed how the results from a survey instrument may be evaluated using a fully Bayesian TRT model. Here, in a single analysis, IRT model parameters were estimated, the testlet structure of the survey instrument was modeled, and the covariates were incorporated. Using the 95% posterior density criterion, the results of the Bayesian analyses suggested that a number of covariates had strong relationship with the latent variable *changeability*. Use of *Tamoxifen*, and *Work Status* (being employed) were both positively associated with changeability while *White* group membership, *Age*, and *Months Since Diagnosis* were negatively related to changeability. Furthermore, the Bayesian findings indicated that *Hispanic* group membership, *Income*,

and *Married* were not significantly related to changeability. The results of post-hoc regression analyses using theta as the dependent variable also agreed with the Bayesian results that use of *Tamoxifen* and *Age* were both related to the latent trait. However, the other covariates that the Bayesian analyses identified as significant—i.e., *Months Since Diagnosis*, *White* group membership and *Work Status*—were not identified as such by the post-hoc approach.

The local dependence problem was also examined for the survey data. The comparison of the point estimates of the theta parameters revealed that they were highly similar for the TRT and IRT models. However, when the impact of not modeling the dependency was examined in the context of precision, it became evident that IRT assumed greater precision than TRT, which resulted in the variances of the IRT theta posteriors being about 50% of their TRT counterparts, when averaged over all respondents. Previous research (e.g., Wang, 2002) suggests that the greater confidence associated with the IRT estimates may be due to the model's failure to account for the testlet dependency.

5.2 Significance of Results

Understanding the variables that are associated with item, person, and testlet parameters is important for both practitioners and researchers. This study utilized a Bayesian TRT model that simultaneously estimates ability and item parameters, models the testlet structure of the test, and incorporates the covariates directly into the model that may help identify some of the potential reasons *why* certain parameter tend to take on certain values. The results of the simulation study showed that the Bayesian model

performs quite similarly to the post-hoc approach for the conditions studied here. The results of both empirical studies also confirm that in a number of cases the two approaches arrive at the same conclusions; however, this is not always the case, as can be seen by the fact that three of the covariates in the PTGI data that were identified as non-significant by the post-hoc approach were identified as significant by the Bayesian approach. On the one hand, this is good news: the Bayesian approach is by-and-large replicating the findings of the post-hoc analysis, and moreover it may be capturing some of the ability-covariate relationships that the conventional post-hoc approach is not capturing. On the other hand, given the expectation that the Bayesian estimates of the coefficients' errors (i.e., posterior variances) should be, if not superior, at least larger than the post-hoc errors, it comes somewhat as a surprise that the Bayesian approach demonstrated more power in detecting relationships than the post-hoc approach. One potential explanation for this paradox was suggested by the simulation results: a possible tendency of the Bayesian approach to inflate coefficient estimates. In any case, the benefit from using this approach, at least under the conditions studied here, seems very small and therefore it may be outweighed by practical concerns such as computational time. Each multiple-chain calibration in this study took approximately 15 hours on a 1.67 GHz processor. Contrast this to a post-hoc regression run, which took under five seconds per replication, it is clear why the Bayesian method may be less preferred. Of course, the Bayesian software is simultaneously estimating the response model parameters along with the covariate slopes, but these parameters could also be obtained from a non-Bayesian calibration software for the purposes of post-hoc analyses.

An important advantage of the TRT model used here is that it models testlet effect. Since this feature of the model was not the primary focus of the study, the simulation study did not include conditions to study the impact of this parameter in the context of covariates. However, the empirical study using the PTGI data investigated the difference in estimated precision when testlet dependency is ignored and found that, when averaged across all examinees, the variance of the posterior of the proficiency parameter was reduced by approximately 50%. If the TRT model is working as intended, this difference is attributable to the failure of the IRT to model testlet dependency. The overall testlet effect (as represented by the variance of the testlet parameter) was modest in four of the five testlet, however for Spiritual Change, it was close to .5. According to Wainer et al. (2007), $\sigma_\gamma^2 = .5$ is plausible for many testing situations and given that it is on the same scale as theta, which has unit variance, .5 can be interpreted as a sizable effect. Overestimation of precision could have important consequences in testing such as premature ending of testing in an adaptive testing context or more generally, it threatens the validity of our score interpretations.

5.3 Limitations and Directions for Future Research

As is the case with all simulation studies, the simulation part of this study cannot be generalized beyond the conditions studied here. The relative performance differences between the Bayesian and post-hoc regression approaches could vary more drastically depending on conditions such as sample size and number of items. For example, the Bayesian approach could outperform post-hoc approach in situations where the sample sizes and item numbers are smaller. As discussed in chapter 3, the Bayesian method used

here estimates the covariate coefficients simultaneously with the model parameters and therefore uses the "extra" information obtained from the item responses in estimating covariate coefficients, and similarly uses the covariate information in estimating model parameters. This sharing of information could be especially useful in small sample situations and give Bayesian approach an advantage and deserves study. As such, examining the performance of the two approaches under different conditions, particularly small-sample and short test length cases, is warranted for future studies.

Another limitation of the simulation study is that only the ability parameter covariate was studied. The model should be tested for the recovery of item and testlet parameter covariates in a simulation study. Study conditions should include the strength of the relationship between the covariates and the model parameters, which should be varied across a range of realistic values, including no relationship.

Lastly, even though the potential benefit of having the testlet parameter is shown using the PTGI data, having an additional parameter means losing a degree of freedom. Therefore, presence of gamma in some instances may result in worse estimates for all model parameters. As always, the prudent researcher should look at model fit before making inferences.

5.4 Conclusion

The findings of this research indicate that the Bayesian approach performs very similarly to the post-hoc approach in estimating covariate relationships. Given the supposed advantages of the Bayesian approach discussed in detail in chapter 2, this outcome was somewhat unexpected. As mentioned in the previous section, other

conditions should be investigated in future studies to better understand the strengths and weaknesses of the model. Theoretically, covariates have the potential inform the *whys* of test development, validity, survey design, and factor structure simultaneously and Bayesian models are likely to play a prominent role in these inquiries in the future. The current computing power is perhaps the biggest limitation of this method, making it unsuitable for some operational uses. Since an average multiple-chain run took about 15 hours under the conditions studied here, simulation research, which typically requires large numbers of replications, faces even greater computing challenges. However, computing power advances rapidly and these practical problems are likely to be short-lived. Given their potential, Bayesian TRT, and fully Bayesian models in general, are likely to continue to attract growing interest.

APPENDIX

THE POSTTRAUMATIC GROWTH INVENTORY

The next set of questions asks about changes in your life after you had your breast cancer. For each of the statements below, please describe the amount of change you experienced.

No change	Very small change	Small change	Moderate change	Great change	Very great change
0	1	2	3	4	5

Factor I: Relating to Others
Knowing that I can count on people in times of trouble.
A sense of closeness with others.
A willingness to express my emotion.
Having compassion for others.
Putting effort into my relationships.
I learned a great deal about how wonderful people are.
I accept needing others

Factor II: New Possibilities
I developed new interests.
I established a new path for my life.
I'm able to do better things with my life.
New opportunities are avail which would not have been otherwise.
I'm more likely to try to change things which need changing.

Factor III: Personal Strength
A feeling of self-reliance.
Knowing I can handle difficulties.
Being able to accept the way things work out.
I discovered that I'm stronger than I thought I was.

Factor IV: Spiritual Change
A better understanding of spiritual matters.
I have a stronger religious faith.

Factor V: Appreciation of Life
My priorities about what is important in life.
An appreciation for the value of my own life.
Appreciating each day.

BIBLIOGRAPHY

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Baldwin, P., Keller, L. A., & Hambleton, R. K. (2004, April). *Using auxiliary information and Bayesian techniques to improve parameter estimation with small samples*. Paper presented at the meeting of the National council on Measurement in Education, San Diego, CA.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F.M. Lord & M.R. Novick, *Statistical theories of mental test scores* (Pps. 392-479). Reading, MA: Addison-Wesley.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika, 37*, 29-51.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for *n* dichotomously scored items. *Psychometrika, 35*, 179-197.

Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.

Chang, S., Plake, B. S., & Ferdous, A. A. (2005, April). *Response times for correct and incorrect item responses on computerized adaptive tests*. Paper presented at the 2005 annual meeting of the American Educational Research Association, Montréal, Canada.

Gelman, A., & Rubin, D. B. (1993). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-472.

Gelman, A., Carlin, J. B., Stern, H. S. & Rubin, D. B. (1995). *Bayesian Data Analysis.* London: Chapman & Hall.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education, 18*, 351-380.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*: Kluwer-Nijhoff.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*: Sage Publications.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *54*, 93-108.

Justice, L. M., Bowles, R. P., & Skibbe, L. E. (2006). Measuring preschool attainment of print-concept knowledge: A study of typical and at risk 3- to 5-year-old children using item response theory. *Language, Speech, and Hearing Services in Schools*, *37*, 224–235,

Keller, L. A. (2002). *Small sample item parameter estimation in the three parameter logistic model: Using collateral information.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Keller, L. A., Swaminathan, H., & Sireci, S. G. (2003). Evaluating Scoring Procedures for Context-Dependent Item Sets. *Applied Measurement in Education*, *16*, 207-222.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. *Applied Psychological Measurement, 25,* 163-176.

Little, H. (2006). Calibrated Bayes: A Bayes/frequentist roadmap. *The American Statistician*, 60, 213-223.

Lord, F. M. (1952). *A theory of test scores.* (Psychometric Monograph No.7) Iowa City, IA: Psychometric Society.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrica*, *47*, 149-174.

MathSoft, I., 1999. S-Plus 2000 Professional Release 2. MathSoft, Inc., Seattle, WA.

Mislevy, R. J., & Bock, R. D. (1983). *BILOG: Item and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Mislevy, R. J., & Bock, R. D. (1985). *BILOG: Item and test scoring with binary logistic models* [computer program]. Mooresville, IN: Scientific Software.

Patz, R. J., & Junker, B. W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24,* 146-178.

Patz, R. J., & Junker, B. W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24,* 342-366.

Samejima, F. (1969). Estimation of latent trait ability using a response pattern of graded scores. *Psychometrika Monograph*, No. 17.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement, 28,* 237-247.

Smith, R. W. (2000). An exploratory analysis of item parameters and characteristics that influence item level response time (Doctoral dissertation, University of Nebraska-Lincoln, 2000). Dissertation Abstracts International, 61, 1812.

Spiegelhalter, D. J., Thomas, A., Best, N. G., & Gilks, W. R. (1997). *BUGS: Bayesian inference using Gibbs sampling* (Version 0.6) [Computer program]. Cambridge, UK: University of Cambridge, Institute of Public Health, Medical Research Council Biostatistics Unit.

SPSS Inc. (2005). SPSS Base 14.0 for Windows User's Guide. SPSS Inc., Chicago IL.

Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7, 175 –191.

Swaminathan, H., & Gifford, J. A. (1985). Bayesian estimation in the two-parameter logistic model. *Psychometrika*, *50*, 349 –364.

Thissen, D. (1991). *MULTILOG: multiple, categorical item analysis and test scoring using Item response theory* (version 7.03) [computer software]. Chicago: Scientific Software.

Thissen, D., Steinberg, L. & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement 26*, 247-260.

Wainer, H. (1995), Precision and differential item functioning on a testlet-based test: The 1991 Law School Admissions Test as an example, *Applied Measurement in Education,* 8(2), 157-187.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet Response Theory*. New York: Cambridge University Press.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University Press.

Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement, 24*, 185-201.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice, 15(1)*, 22-29.

Wang, X., Bradlow, E.T., and Wainer, H. (2002). A General Bayesian Model for Testlets: Theory and Applications. *Applied Psychological Measurement*, *26*, 109-128.

Wang, X., Bradlow, E. T., & Wainer, H. (2004). SCORIGHT: A computer program for scoring tests built of testlets including a model for covariate analysis (Version 3.0) [Computer program]. Princeton, NJ: Educational Testing Service.

Wang, X., Bradlow, E. T., & Wainer, H. (2005). User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module.

Wollack, J.A., Bolt, D. M., Cohen, A. S., & Lee, Young-Sun. (2002). Recovery of item parameters in the nominal response model: A comparison of Marginal Maximum Likelihood estimation and Markov chain Monte Carlo estimation. *Applied Psychological Measurement, 26,* 339-352.

Zenisky, A.L., Hambleton, R. K., & Sireci, S.G. (2002). Identification and evaluation of local item dependencies in the Medical College Admission Test. *Journal of Educational Measurement, 39,* 291-309.

3474-19